

# 가가가 가? 시리야 이거 표준어로 번역해줘

Audio 딥러닝

붉은악마의자식 🐱 권보영 서지영 한채현

# Project overview



# Motivation

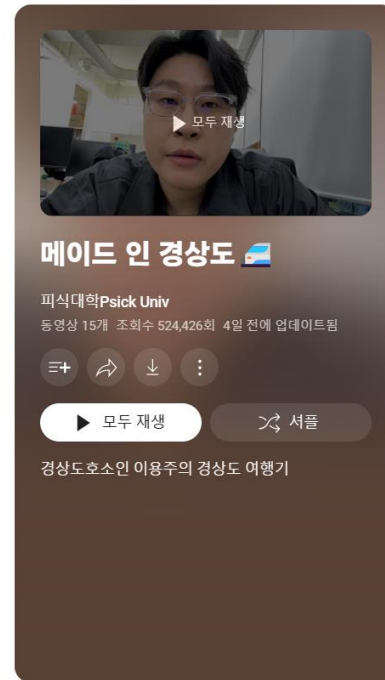
## 내 남편과 결혼해줘 사투리 ▼



국밥씬 편안해지는 사투리 더빙 #shorts #내남편과...



이상한 사투리 연기 #박민영 #이기광 #사투리연기



1. 경상도 호소인 논란? 할 말은 해볼라예  
피식대학Psick Univ · 조회수 116만회 · 4개월 전
2. 경상도 호소인? "두고봐뿌아지예"  
피식대학Psick Univ · 조회수 109만회 · 3개월 전
3. 경상도 호소인? 유물 다 캐키웠예 [경북 경주1]  
피식대학Psick Univ · 조회수 83만회 · 3개월 전
4. 세계에서 4번째로 높은 놀이기구 깔끼고 왔습니데이 [경북 경주2]  
피식대학Psick Univ · 조회수 64만회 · 3개월 전
5. 경상도 호소인? 드디어 내 고향 부산에 와칸다쓰예 [부산1]  
피식대학Psick Univ · 조회수 92만회 · 2개월 전
6. 경상도 호소인? 부산 야구의 성전으로 입성해웠예 [부산2]  
피식대학Psick Univ · 조회수 53만회 · 2개월 전

## Title

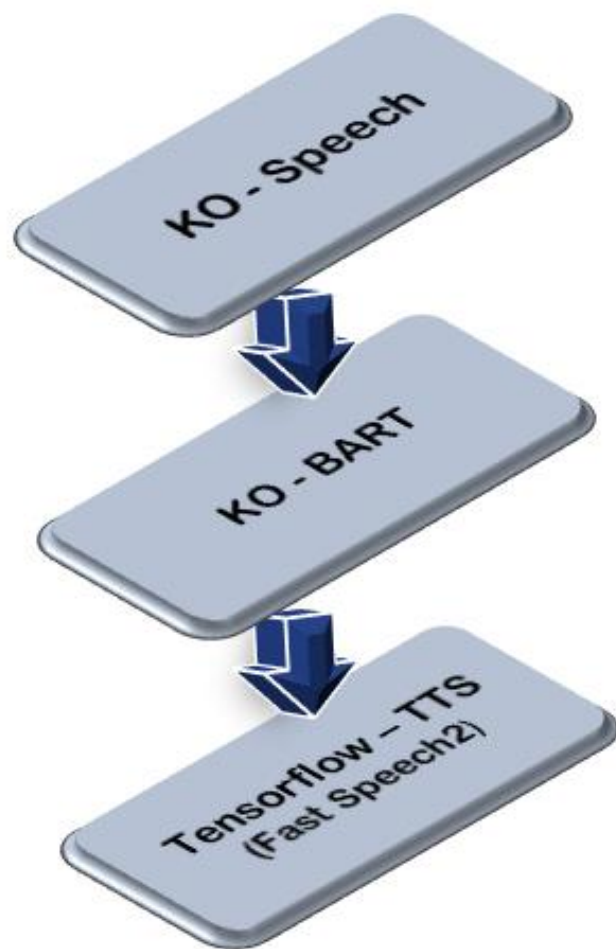


## 사투리 인식 및 표준어 변환 시스템

## Workflow

사투리 발화 → Speech to Text → 텍스트 → Text to Speech → 표준어 발화

## Process



STT (Speech-to-Text)  
Ko-Speech



Translation  
(Text-to-Text)  
Ko-BART




TTS (Text-to-Speech)  
Tensorflow TTS

**Main content**



# Data

## 한국어 방언 발화(경상도)



경상도

#지능형플랫폼 구축 # AI 돌봄 서비스 # 스마트시티 데이터 허브

### 한국어 방언 발화(경상도)

분야 한국어 유형 오디오, 텍스트

구축년도 : 2020 갱신년월 : 2023-03 조회수 : 6,784 다운로드 : 1,271  
용량 : 322.98 GB

- 경상도 지역 화자가 발화한 3000시간 이상의 음성 데이터 &
- 방언 텍스트와 그에 대응하는 표준어 대응쌍
- JSON 포맷의 데이터 파일로 구성

```

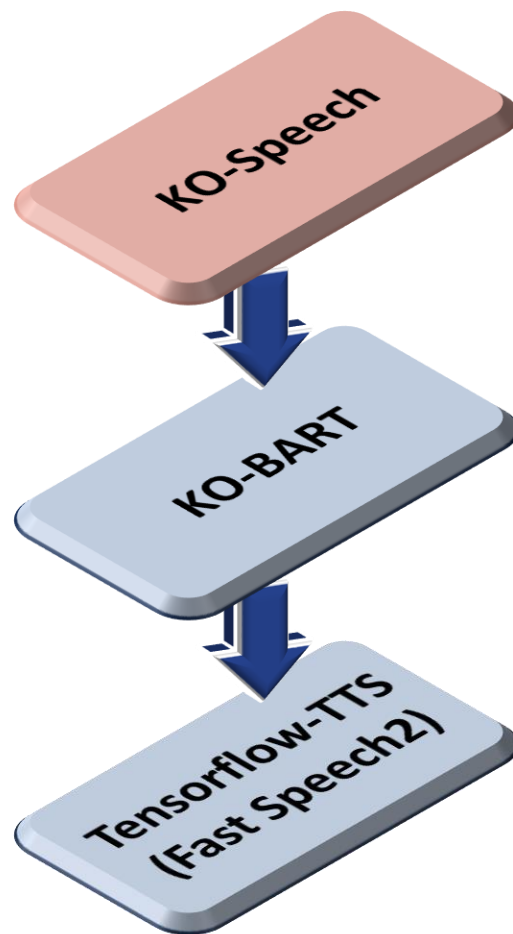
DKCI20000001.json
스키마: <선택된 스키마가 없음>
19 {
20   "id": "1",
21   "name": "null",
22   "age": "50대",
23   "occupation": "사무 종사자",
24   "sex": "남성",
25   "birthplace": "경남",
26   "principal_residence": "부산",
27   "current_residence": "경남",
28   "education": "고졸"
29 },
30 {
31   "id": "2",
32   "name": "null",
33   "age": "50대",
34   "occupation": "전문가 및 관련 종사자",
35   "sex": "여성",
36   "birthplace": "부산",
37   "principal_residence": "경남",
38   "current_residence": "경남",
39   "education": "대졸"
40 },
41 ],
42 "setting": {
43   "relation": "부부"
44 },
45 "utterance": [
46   {
47     "id": "DKCI20000001.1.1.1",
48     "form": "자 음식을 멀 좋아하느냐 하면은",
49     "standard_form": "자 음식을 멀 좋아하느냐 하면은",
50     "dialect_form": "자 음식을 멀 좋아하느냐 하면은",
51     "speaker_id": "1",
52     "start": 0.01,
53     "end": 3.45,
54     "note": "",
55     "eojeolList": [
56       {
57         "id": 1,
58         "eojeol": "자",
59         "standard": "자",
60         "isDialect": false
61       },
62     ]
  }
]
  
```

사투리 품 &  
표준어 품

음성녹음에서 문장이 시작 및 끝나는 시간



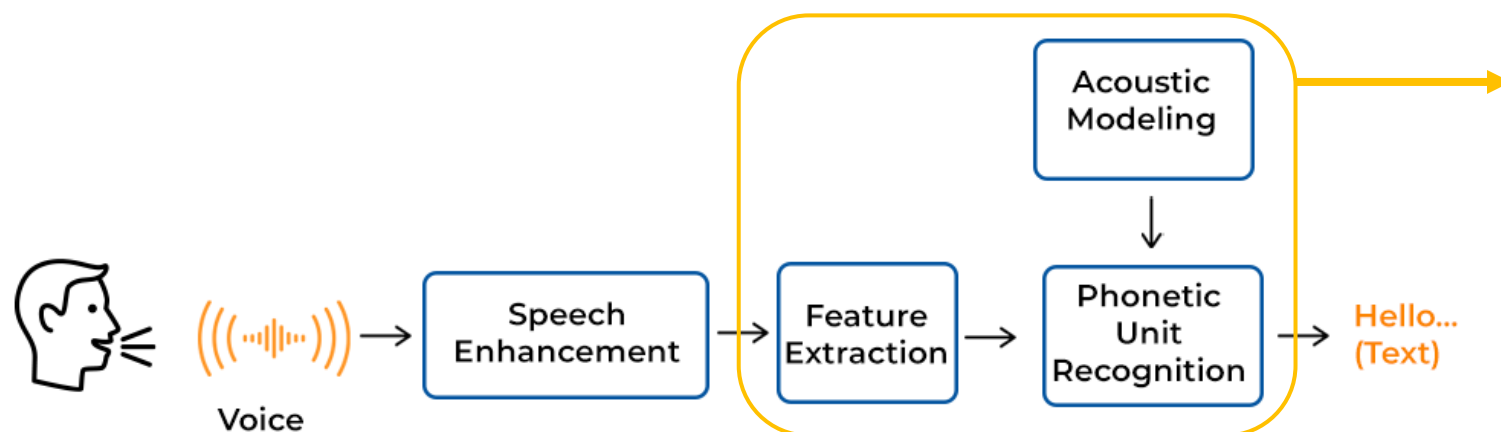
## Process



# Methodology

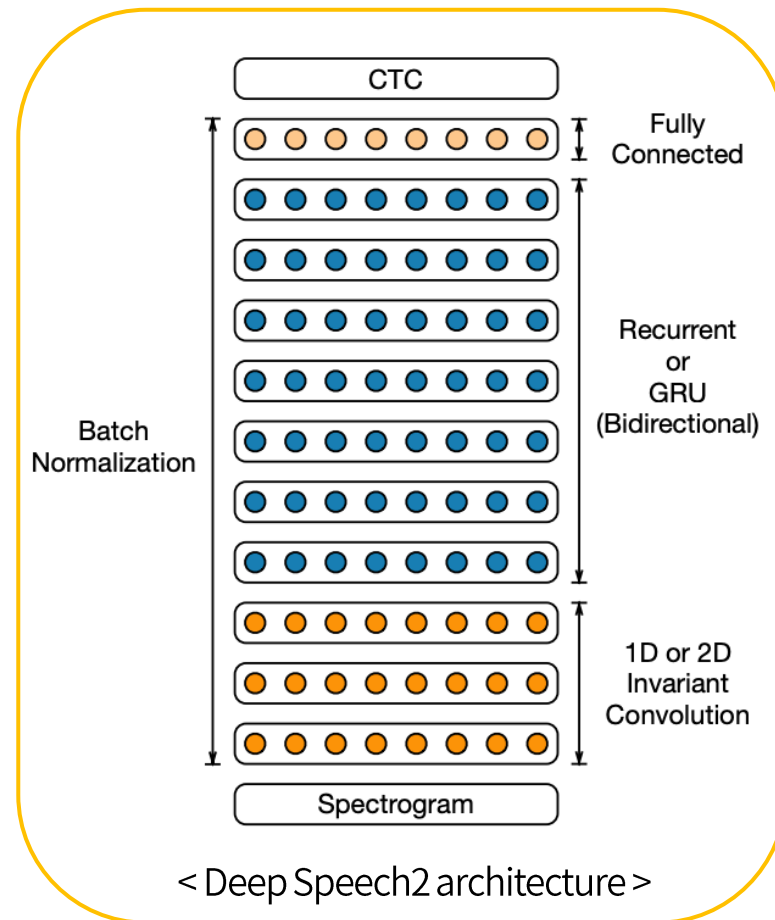
Ko-Speech (DeepSpeech2)

## SPEECH RECOGNITION PROCESS

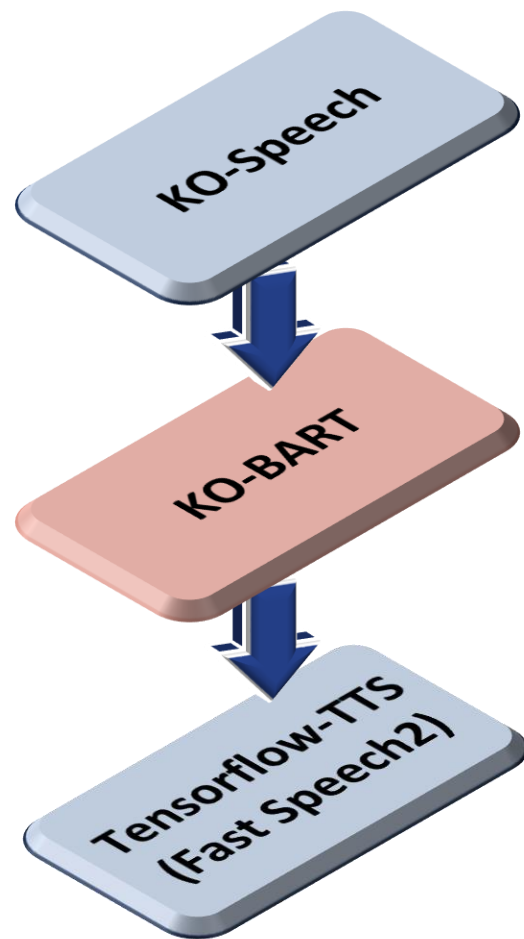


**Ko-Speech:** Pytorch 기반의 한국어만 지원하는 모델이며 DeepSpeech2, Speech Transformer 등 다양한 백본 구조를 갖고 있다.

사투리 발화 → 텍스트  
음성인식 모델로 Ko-Speech 선정!

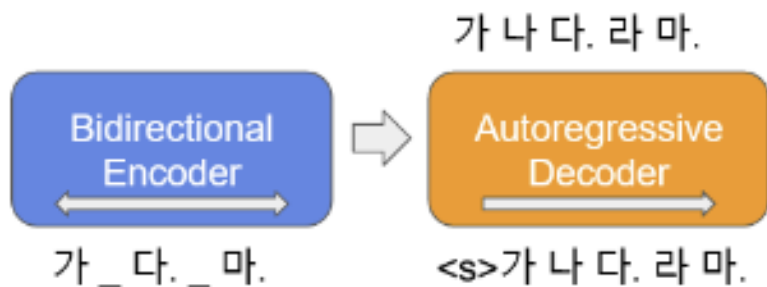
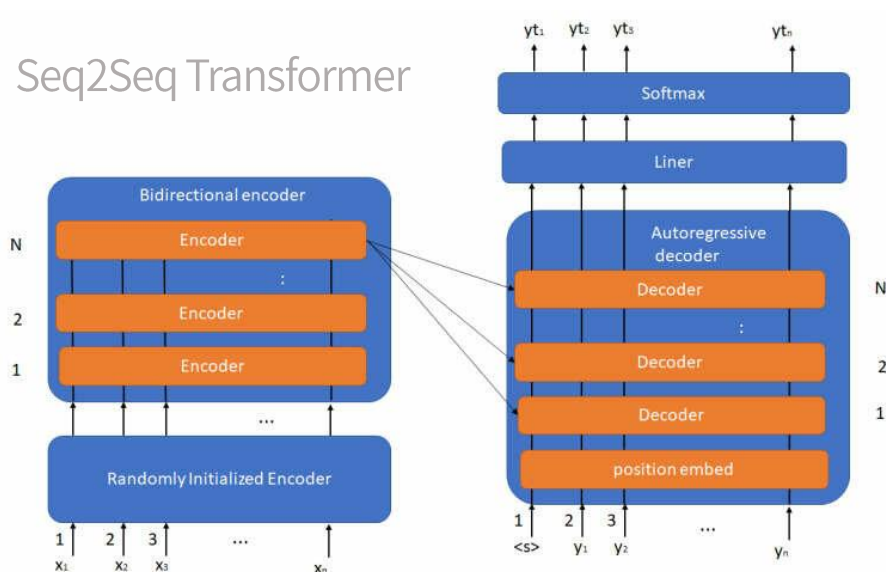


## Process



# Methodology

## Ko-Bart



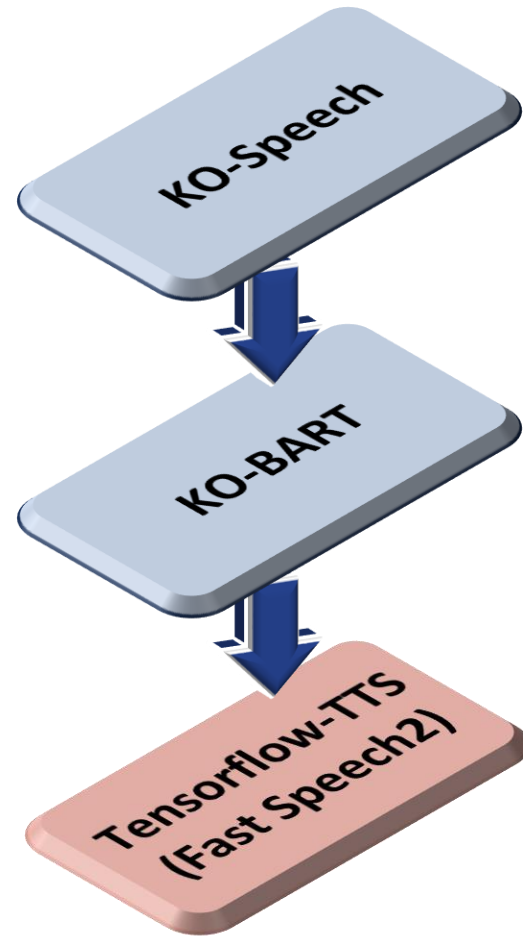
**BART**(Bidirectional Auto Regressive Transformer)  
= BERT + GPT

**Ko-Bart**: 한국어 BART로, 40GB 이상의 한국어 텍스트에  
대해서 학습한 한국어 encoder-decoder 언어 모델

사투리 → 표준어

Text2Text(Translation) 모델로 Ko-BART 선정 !

## Process



## Methodology

### TensorFlow TTS



#### TensorSpeech/ TensorFlowTTS

🤖 TensorFlowTTS: Real-Time State-of-the-art Speech Synthesis for Tensorflow 2 (supported including English, French, Korean, Chinese, German and Easy to adap...

Ak 32  
Contributors

🔗 32  
Used by

☆ 4k  
Stars

🍴 787  
Forks



TensorFlow2 기반 Tacotron-2, Melgan, Multiband-Melgan, FastSpeech, FastSpeech2와 같은 실시간 음성 합성 architecture 제공

표준어 텍스트 ➡ 표준어 발화

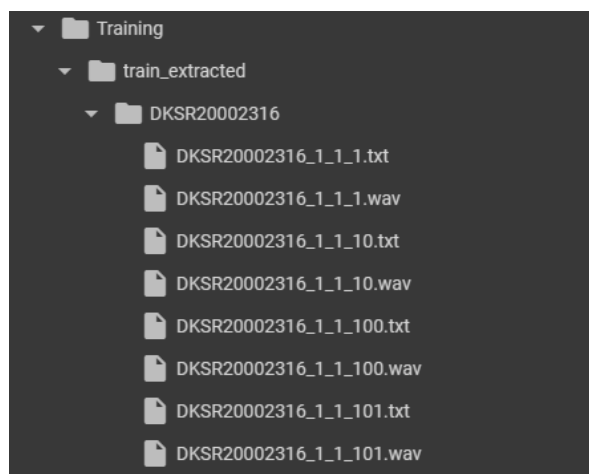
Text-to-Speech 모델로 TensorFlowTTS 선정 !

# Pre-Processing

## Ko-Speech

### 1. 데이터 폴더 생성

오디오(.wav)-텍스트(.txt) pair을 생성



Our data (ex)

### 2. 전사 파일(transcript.txt) & 단어 사전(labels.csv) 생성

벡터화된 전사 텍스트 포함하는 transcript.txt 생성

Character 수준에서 단어 빈도수 측정하는 labels.csv 생성

```

5/DKSR20002415_1_1_18.wav 라디오가 그 30 244 87 8 3 6
5/DKSR20002415_1_1_19.wav 대학생들 연합 동아리식으로 하는 건데 54 111 44 32 3 138 347 3 108 18 42 13
5/DKSR20002415_1_1_20.wav 저는 다양한 주제를 다루고 있어요 최근에 한 게 48 5 3 17 273 35 3 63 24
5/DKSR20002415_1_1_21.wav 그 스트레스 편이라고 6 3 61 123 282 61 3 160 4 30 7
5/DKSR20002415_1_1_22.wav 스트레스받은 게스트들을 모셔가지고 61 123 282 61 185 23 3 14 61 123 32 2
5/DKSR20002415_1_1_23.wav 그 게스트분들이 뭐 아나운서를 준비하거나 6 3 14 61 123 109 32 4 3 29 3 1
5/DKSR20002415_1_1_24.wav 아니면 또 유튜브나 이런 데 방송하는 친구들이거든요 18 28 26 3 82 3 139 386 2
5/DKSR20002415_1_1_25.wav 그래서 그분들이 약간 스트레스에 어떤 식으로 받았는지 6 41 10 3 6 109 32
5/DKSR20002415_1_1_26.wav 약간 그런 거를 방송을 통해 얘기하고 약간 프로그램 95 71 3 6 31 3 11 38 3 18
5/DKSR20002415_1_1_27.wav 프로그램을 통해 스트레스를 해소해주는 그런 방송을 해봤어요 245 36 6 417
5/DKSR20002415_1_1_28.wav 그리고 내일 녹음하는 것도 있는데 6 42 7 3 65 55 3 755 92 9 5 3 68 16 3
5/DKSR20002415_1_1_29.wav 내일은 또 상상 특집이라고 내가 상상해서 어떤 거를 하고 싶은가 65 55 23 3 8
5/DKSR20002415_1_1_30.wav 약간 그런 식으로 할려는 예정인데 95 71 3 6 31 3 135 45 36 3 98 116 5 3
5/DKSR20002415_1_1_31.wav 제가 앞서 앞 전에 게스트가 나 아 디제이가 나갔다고 해서 24 8 3 351 10 3 35
5/DKSR20002415_1_1_32.wav 너무 진행하기 힘들었는데 106 70 3 84 196 9 27 3 250 32 58 58 5 20
5/DKSR20002415_1_1_33.wav 또 저희 프로그램 짐 라디오의 연합 동아리가 프로그램을 다섯 개하는데 82 3 4
5/DKSR20002415_1_1_34.wav 저희 프로그램만 지금 게스트를 받고 있거든요 48 241 3 245 36 6 417 51 3 13 8
5/DKSR20002415_1_1_35.wav 근데 디제이가 나감으로써 게스트들이 계속 도와주다 보니까 79 20 3 244 24 4 8
5/DKSR20002415_1_1_36.wav 프로그램도 점점 재밌어지고 245 36 6 417 16 3 200 200 3 159 246 15 13 7

```

전사파일 예시

	A	B	C
1	id	char	freq
2		0 <pad>	0
3		1 <sos>	0
4		2 <eos>	0
5		3	1929201
6		4 이	235990
7		5 는	173824
8		6 그	172876
9		7 고	156106
10		8 가	127396
11		9 하	98720
12		10 서	89151
13		11 거	87749
14		12 예	86445
15		13 지	86427
16		14 게	83571
17		15 어	83226
18		16 도	80061
19		17 다	79712

단어사전 예시

# Pre-Processing

## Ko-BART

### 1. 특정 문자열 제거

- 괄호 및 중복 문자 제거

```
step1 = re.compile(r'&#w+&')
step2 = re.compile(r'&#(+)')
step3 = re.compile(r'&#{[^\]}*')
step4 = re.compile(r'&#-([^\]})*-')
```

Text : &name2& 님은 뭐~ {laughing} (() 힘들 때 -위- 위로가 되어 준 노래 같은 게 있으신가요?  
 Step 1 : 님은 뭐~ {laughing} (() 힘들 때 -위- 위로가 되어 준 노래 같은 게 있으신가요?  
 Step 2 : 님은 뭐~ {laughing} 힘들 때 -위- 위로가 되어 준 노래 같은 게 있으신가요?  
 Step 3 : 님은 뭐~ 힘들 때 -위- 위로가 되어 준 노래 같은 게 있으신가요?  
 Step 4 : 님은 뭐~ 힘들 때 위로가 되어 준 노래 같은 게 있으신가요?

Text : 어~ {반가움} &name123&! (() 반갑고 ㅋㅋ -오- 오랜만이다  
 Step 1 : 어~ {반가움} ! (() 반갑고 ㅋㅋ -오- 오랜만이다  
 Step 2 : 어~ {반가움} ! 반갑고 ㅋㅋ -오- 오랜만이다  
 Step 3 : 어~ ! 반갑고 ㅋㅋ -오- 오랜만이다  
 Step 4 : 어~ ! 반갑고 ㅋㅋ 오랜만이다

### 2. 데이터 정제

- 결측값 포함 행 제거
- make\_cleaned\_text

```
def make_cleaned_text(text):
    after_step1=step1.sub('',text)
    after_step2=step2.sub('',after_step1)
    after_step3=step3.sub('',after_step2)
    after_step4=step4.sub('',after_step3)

    return after_step4

for idx, row in df.iterrows():
    df.at[idx, '사투리'] = make_cleaned_text(row['사투리'])
    df.at[idx, '표준어'] = make_cleaned_text(row['표준어'])

df = df.iloc[:, 1:]

df.to_csv('data_fixed.tsv', index=False, sep='\t')
}
```

### data\_fixed.tsv

	사투리	표준어
0	저는 보는 거 이케	저는 보는 거 이렇게
1	좋아하고 이케 고양이 만지는데	좋아하고 이렇게 고양이는 만지는데
2	그 이케 강아지를 잘 못 안겠더라고요 그 물	그 이렇게 강아지를 잘 못 안겠더라고요 그 물
3	이케 사람도 물컹하긴 하지만	이렇게 사람도 물컹하긴 하지만
4	개들이 좀 더 이케 잡기가 저는 그	개들이 조금 더 이렇게 잡기가 저는 그



# Experiments



## Experiment 1: Ko-Speech

### ✓ Setting 1

- Dataset : Kspon (한국어 음성 발화)
- Num\_epochs : 1.5
  - 16.3시간 소요
- Train/Valid Split : 600000 / 22500

### ▪ Setting 2

- Dataset : DKSR (한국어 방언 발화)
- Num\_epochs : 1.0
- Train/Valid Split : 8:2 (210000:51064)



두 가지 setting 모두 실험 결과,  
KSPON의 학습 결과(*Settings 1*)가  
DKSR의 inference에서 성능이 높았습니다.

## Evaluation Metric

CER (Character Error Rate)

인식된 문자열과 정답 문자열 사이의  
문자 오류 비율을 나타내는 지표

$$CER = \frac{S + D + I}{N}$$

I: minimum number of insertions

S: substitutions

D: deletions

N: total number of characters

## Experiment 2 : Ko-BART

- Setting
  - Num\_epochs : 3
  - 2 hours

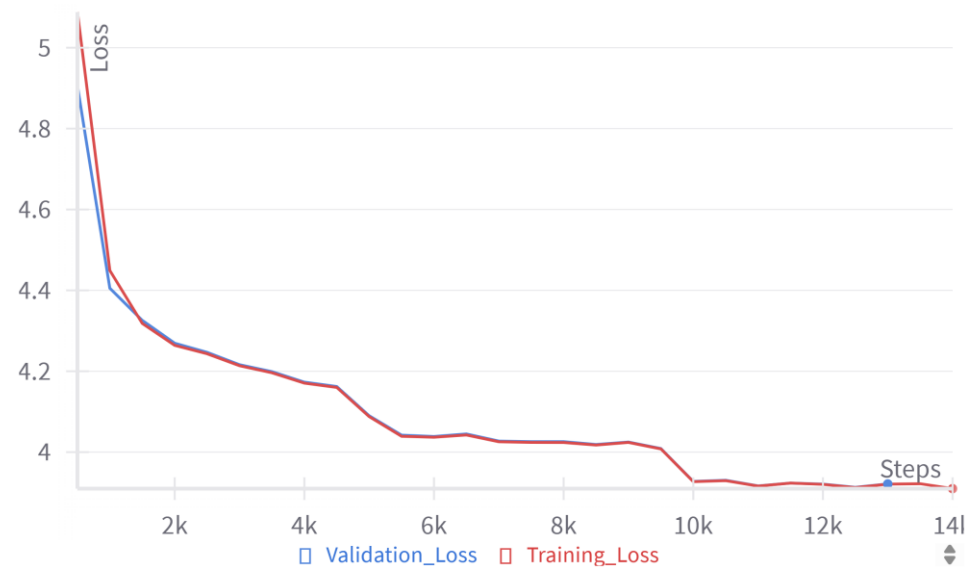


```
model = pipeline(
    'text2text-generation',
    model='heegyu/kobart-text-style-transfer'
)

def transfer_text_style(text, target_style, **kwargs):
    input = f"{target_style} 말투로 변환:{text}"
    out = model(input, max_length=64, **kwargs)
    print(text, target_style, out[0]['generated_text'], sep="->")
```

[heegyu/kobart-text-style-transfer · Hugging Face](#)

Ko-BART Loss Curve

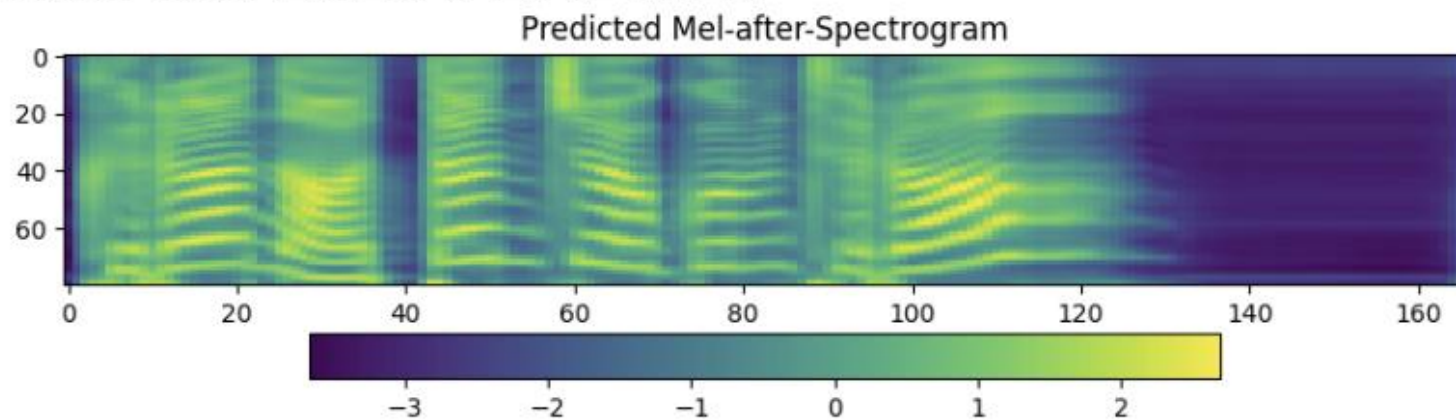


## Experiment 3 : Tensorflow-TTS

### FastSpeech2 불러오기

Ko-Speech 결과(음성인식) : 그런 어떤 좀 들어?

Translated Standard (표준어로 출력): 그런 어떤 조금 들어?



▶ 0:00 / 0:01 🔊 ⋮

## Evaluation : Ko-Speech

### Valid dataset 250 files

Ncor: 613   Nsub: 1005   Ndel: 168   Nins: 78  
 Lev distance: 1733   length: 4563

➔ CER: 0.379793

- good example

79) [DKSR20000890\_1\_1\_86.wav]  
 [transcript] 어 어떤 사람 만났을 때 어 그런 생각이 들면은  
 [prediction] 어 어떤 사람 만났을 때 어 그런 생각이 들면은  
 Ncor: 9   Nsub: 0   Ndel: 0   Nins: 0  
 Lev distance: 0   length: 18  
 WER: 0.0   CER: 0.0

323) [DKSR20000891\_1\_1\_88.wav]  
 [transcript] 일곱 시까지 항상 만화를 많이 했잖아  
 [prediction] 일곱시까지 항상 만화를 많이 했잖아.  
 Ncor: 3   Nsub: 2   Ndel: 1   Nins: 0  
 Lev distance: 1   length: 15  
 WER: 0.5   CER: 0.06666666666666667

- bad example

196) [DKSR20000890\_1\_1\_206.wav]  
 [transcript] 마냥 솔직한 게 좋지 않냐  
 [prediction] 맞 솔직하매 높지냐.  
 Ncor: 0   Nsub: 3   Ndel: 2   Nins: 0  
 Lev distance: 7   length: 10  
 WER: 1.0   CER: 0.7

222) [DKSR20000890\_1\_1\_232.wav]  
 [transcript] 그러면은 수입이 계속 이렇게 있는 있으신 거잖아요  
 [prediction] 그러면에 쓰비 계속 이렇게 임메 이 주신 받다마야.  
 Ncor: 2   Nsub: 5   Ndel: 0   Nins: 1  
 Lev distance: 13   length: 21  
 WER: 0.8571428571428571   CER: 0.6190476190476191

## Evaluation : Ko-BART

- Inference

```
Write 'q' to exit
Dialect to translate(입력받을 사투리) : 뭐라카노
Translated Standard (표준어로 출력): 뭐라고 하니
Dialect to translate(입력받을 사투리) : 코로나 땀에 쫄 그렇긴했었지
Translated Standard (표준어로 출력): 코로나 땀에 조금 그렇긴했었지
Dialect to translate(입력받을 사투리) : 상황이 근까 쫄 안 좋았다
Translated Standard (표준어로 출력): 상황이 그러니까 조금 안 좋았다
Dialect to translate(입력받을 사투리) : 우리 일상에 엄청 흔한 요소 중에 하나다 아이가
Translated Standard (표준어로 출력): 우리 일상에 엄청 흔한 요소 중에 하나지 않니
Dialect to translate(입력받을 사투리) : 따셨다 아이가
Translated Standard (표준어로 출력): 따셨지 않니
Dialect to translate(입력받을 사투리) : 그까지 해야 우리가 살 수 있는 기라
Translated Standard (표준어로 출력): 거기까지 해야 우리가 살 수 있는 거야
Dialect to translate(입력받을 사투리) : 
```

사투리 → 표준어 변환 Good

## Evaluation : Final Model

입력할 음성 파일

```
# 파이프라인 연결 (Ko-Speech ver.)
import time
start_time = time.time()

src_text = !python3 /content/drive/MyDrive/kospeech/bin/inference.py --model_path '$model.pt' --
audio_path '$input.wav' --device "cpu"
src_text = str(src_text)[2:-2]
print("Ko-Speech Result (음성 인식) : ", src_text)

target_text_ko = generate_text(nlp_pipeline,src_text,num_return_sequences=1,max_length=64)[0]
print(f"Translated Standard Result (표준어체로 출력): {target_text_ko}")

mels, audios = do_synthesis(target_text_ko, fastspeech2, mb_melgan, "FASTSPEECH2", "MB-MELGAN")
print("최종 출력 표준어 발화 : ")
visualize_mel_spectrogram(mels[0])
ipd.Audio(audios, rate=22050)
print("Inference time : ",round(time.time() - start_time, 3),"sec")
```

1. KO-SPEECH

2. KO-BART

3. FAST-SPEECH 2



# Conclusion



## Result

[DKSR20002401\_1\_1\_86.wav]

[ Input ]                    [transcript] 당연히 의리 그리해야 내는 거 맨지로 언제 그리까 하는데



[ Output ]                    [prediction] 당연히 의리 그리 해야 나는 거 처럼 언제 그럴까 하는데



[DKSR20000890\_1\_1\_166.wav]

[ Input ]                    [transcript] 사실 상대방이 얼마나 친절하구 날 좋아해 주냐 그게 더 중해가지구



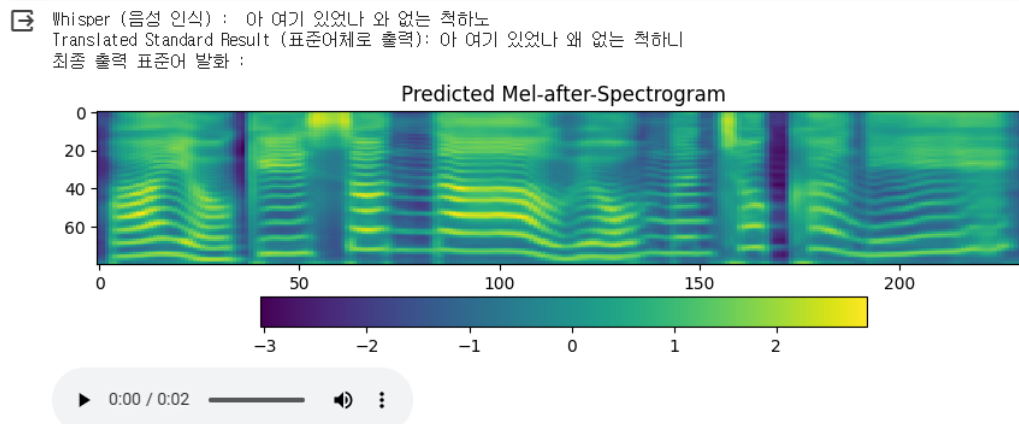
[ Output ]                    [prediction] 사실 단담방 얼마나 친절하고 날 좋아해지냐 그게 더 중해가지지구



# Discussion

## 1. Speech Recognition

- Whisper



## 2. Fine Tuning

## 3. Evaluation Metric

(Whisper version)

input

아 여기 있었나 와 없는 척하노

output

아 여기 있었나 왜 없는 척하니

input

내 좋아했나? 그라운 그때 와 그란건데?

output

나 좋아했나? 그러면 그때 왜 그런건데?

**Thank you**