

# 우리 모두의 블루스

ENGLISH TEXT-TO-JEJU SPEECH READER

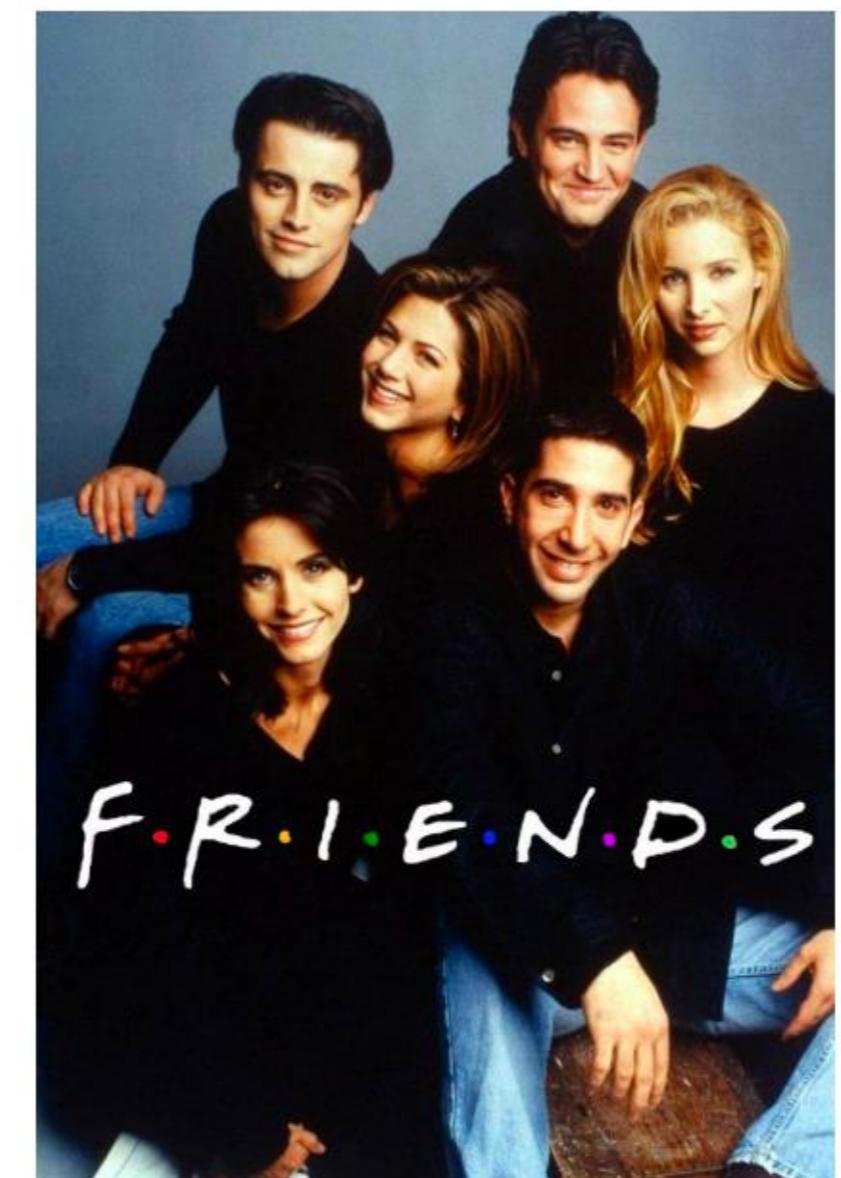
영어 컨텐츠를 제주의 언어로 번역하여 직접 읽어주는 서비스

투빅스 16기 김건우, 장준원, 전민진, 정수연, 이승주

투빅스 17기 임수진, 홍종현



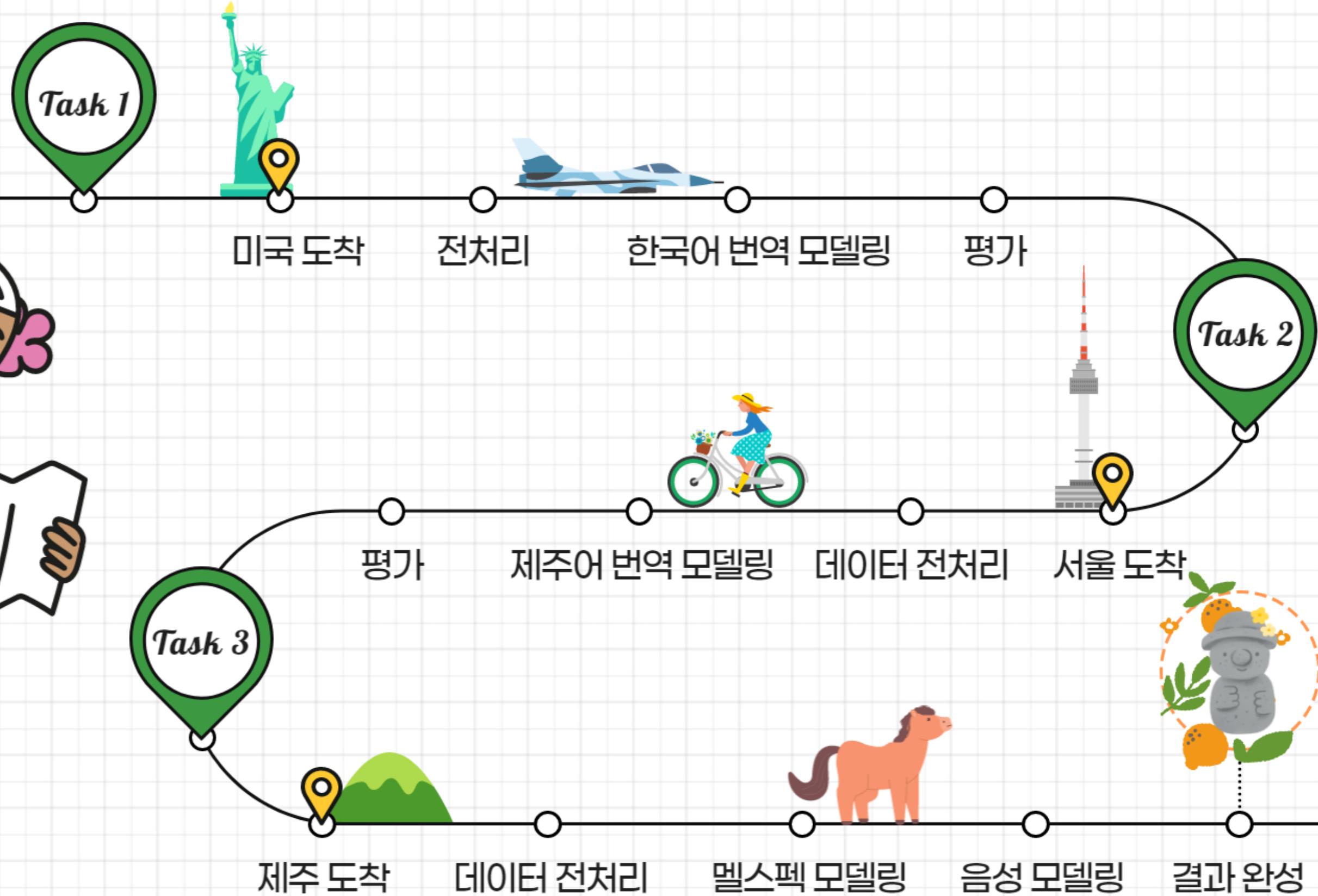
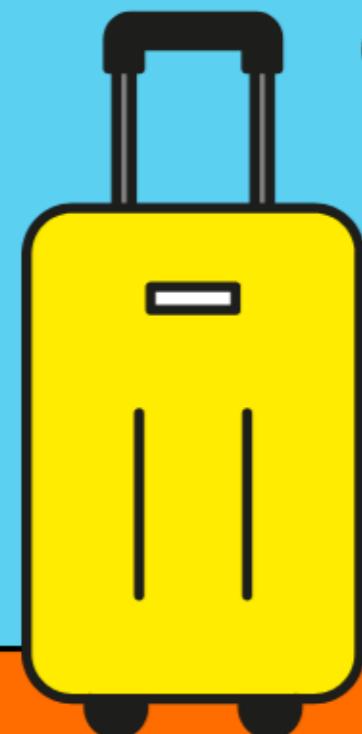
# INTRODUCTION



외국인이 제주도 사투리로 말하면 어떨까?

# PROCESS

미국-제주 여행 코스





## *So, Where we go?*

# 영어-한국어 번역

### Hello World ! , 반갑다 세상아 !

영어가 제주도 사투리로 알맞게 변환 되기 위해서, 먼저 영어 텍스트가 한국어로 알맞게 번역되어야 한다. 이질적인 두 언어가 제대로 번역되기 위해서는 좋은 번역 모델이 필요하다. 대규모 말뭉치(corpus)를 사전 학습한 언어 모델을 사용하면 번역에서 좋은 성능을 낸다. 우리가 구할 수 있는 데이터의 양과 시간을 고려했을 때, pre-train 언어 모델을 사용하는 것이 효과적이므로 mBART를 사용했다. 구축된 대부분의 번역은 문어체를 중점적으로 다룬다. 그러나 우리는 사투리 대화체로의 번역을 중점적으로 다루므로, AI-HUB 구어체 데이터와 Netflix 영화/드라마 자막을 fine-tuning하여 영-한 번역기를 생성했다



# **DATASET**

—  
데이터셋

**AI hub** 영어-한국어  
병렬 코퍼스

AI 번역 엔진 개발을 목적으로 구축한

영어-한국어 번역 말뭉치

뉴스(80만), 정부/지자체 홈페이지,

간행물(10만), 행정 규칙, 자치법규(10만), 한국

문화(10만 문장), **구어체(40만)**, **대화체(10만)**



**Netflix** 영어-한국어  
자막 코퍼스

Netflix 일상 영화 및 드라마 37편 자막 추출

(노트북, 타이타닉, 트루먼쇼, 등)

더 많은 대화체와 구어체 데이터를 확보하기 위

해 LLN(Language Learning with Netflix)

를 통해 영어-한국어 병렬 코퍼스 추출



	Total	Train	Valid
AI hub (En-Kr)	499,991	399,993	99,998
Netflix (En-Kr)	104,161	83,329	20,832



# ENGLISH TO KOREAN TRANSLATION

## Data Preprocessing

### STEP 01 : 지시어 제거 및 소문자 변환

- 대괄호와 소괄호 안에 있는 문자 제거 ( imitating machine ) : you have two messages.  
to record your message, begin speaking at the tone. ( tone )

### STEP 02 : 불필요한 문자 제거

- 구두점 ('.' , ',' , '!' , '?' , ';' , ':')과 영어, 숫자, 한국어를 제외한 모든 문자 제거

### STEP 03 : 중복되는 문장 제거

- 한글 자막의 타이밍이 겹칠경우 다음 자막과 함께 중복으로 추출

i'm sorry, okay?

미안해요, 알았죠? 전부 없었던 일로 해요

i hope that we can just forget the whole thing.

미안해요, 알았죠? 전부 없었던 일로 해요 끊을게요



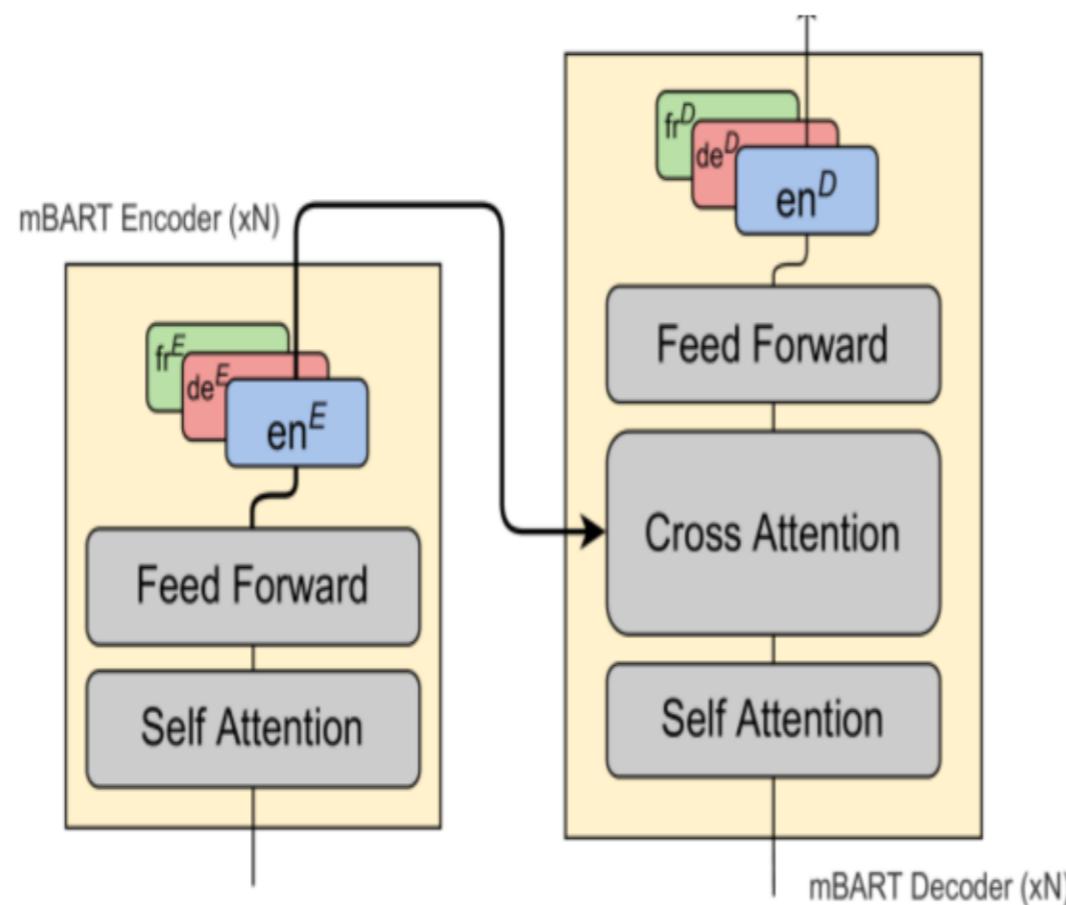
i'm sorry, okay? i hope that we can just forget the whole thing. | 미안해요, 알았죠? 전부 없었던 일로 해요 끊을게요

- 현재 문장이 이전 문장에 포함되어 있다면 현재 문장의 영어 문장만 이전 문장에 추가하고 현재 문장을 삭제
- 이전 문장이 현재 문장에 포함되어 있다면 이전 문장의 영어 문장만 현재 문장에 추가하고 이전 문장을 삭제

# MODEL

**mBART**

Pre-trained on large-scale monolingual based on BART  
English-Korean Neural Machine Translation



- 12 Encoder and Decoder blocks
- 1024 hidden units across 16 attention heads (~ 680M parameters)
- additional layer-normalization layer on top of both the encoder and decoder

## Pretraining Objective Function

Text Infilling

A B C . D E .      A \_ . D \_ E .

Sentence Permutation

A B C . D E .      D E . A B C .

$$\mathcal{L}_\theta = \sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{X \in \mathcal{D}_i} \log P(X|g(X); \theta)$$

# ENGLISH TO KOREAN TRANSLATION

**mBART**

## mBART-large-cc25

- 25개 국어로 pre-train된 model
- Vocabulary size : 50265

## Tokenization

- Pre-trained SentencePiece를 통해 토큰화
- 각 문장의 끝에 eos/sep token '</s>' 추가
- Input 텍스트가 어떤 언어인지 알려주는 '<LID>' symbol token 추가

## BLEU score

- BLEU scoring of generated translations against reference translations
- \* BLEU Score : n-gram에 기반하여 Machine Translation 결과와 target01 얼마나 유사한지 비교하는 성능 측정 방법

-> BLEU Score : 15.55

## mBART Fine-tuning

- optimizer AdamW
- lr 0.0002
- dropout 0.3
- weight-decay 0.0002

...

# ENGLISH TO KOREAN TRANSLATION

**Result**

영어	한국어
To whom do they belong?	그들은 누구에게 속해 있습니까?
You own the stars?	당신은 그 별들을 소유하고 있습니까?
Is that all that is necessary?	그게 전부 필요한 건가요?
Now, how are we going to communicate this?	어떻게 소통할 거야?
And, what's wrong with their user interfaces?	그리고 그들의 사용자 인터페이스는 왜 그래요?

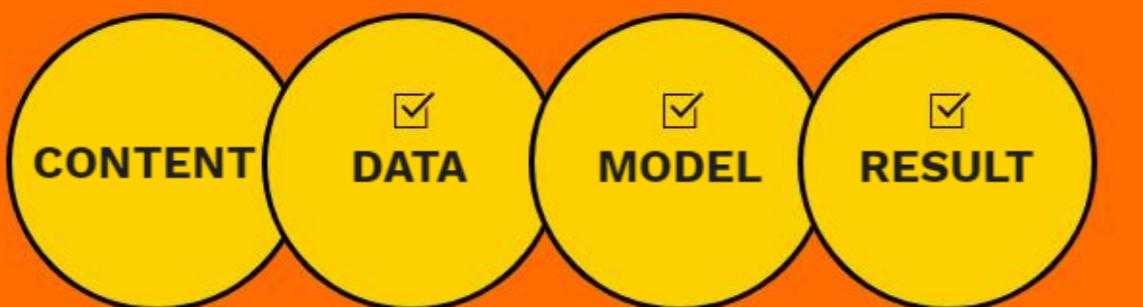


*So, Where we go?*

# 표준어-제주 사투리 번역

## 사라져가는 제주어의 위기, “알고 있수꽈?”

제주어는 제주도에서 사용되는 소수 언어로 2010년 유네스코에 의해 멸종위급 종으로 분류되었다. 현재 딥러닝에서는 attention기법들이 발전해있고 이를 통해 데이터가 적은 순간에도 구문론적, 형태론적, 의미론적으로 좋은 성과를 낸다. 이런 도전에 영감을 받아 제주도 방언 인터뷰 스크립트 기반의 제주어 - 표준어 병렬 corpus인 JIT 데이터셋과 AIHUB 데이터셋으로 신경망 기계 번역 모델을 학습시켜 제주어 - 표준어 병렬 데이터셋을 발전시키고자 한다.



# **DATASET** — 데이터셋

## **AI hub** 한국어 방언 발화 데이터

표준어 텍스트 및 방언 특성을 고려하여 전사한

텍스트 (강원도, 경상도, 전라도, 충청도, **제주도**)

- "form": "(경)/(그렇게)하더라고"
- "standard\_form": "그렇게하더라고"
- "dialect\_form": "경하더라고"



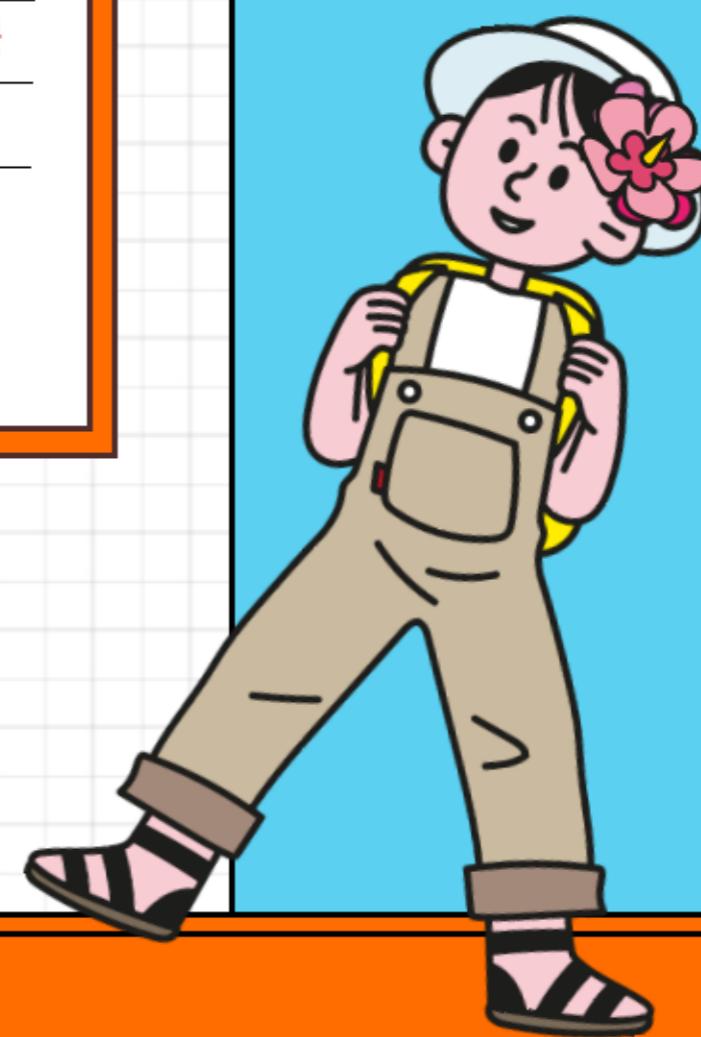
## **Kakao** 제주도 방언 번역 데이터

Kakao Jejueo Datasets for Machine  
Translation and Speech Synthesis

카카오 브레인은 제주도 방언 인터뷰 스크립트를  
통해 **170,000개 이상의 문장을 만들어 제주-한국**  
**병렬 말뭉치**인 JIT(Jejuo Interview  
Transcripts) 데이터 세트를 개발.

# kakao**brain**

	Total	Train	Dev	Test
AI hub (JEJU)	3,216,310	2,547,898	318,487	318,488
JIT	170,356	160,356	5,000	5,000



## Byte Pair Encoding (BPE)

서브워드 분리(subword segmentation) 알고리즘으로 데이터에서  
가장 많이 등장한 문자열을 병합해 데이터를 압축하는 기법.

### - BPE를 활용한 토큰화 절차

① 어휘 집합 구축 : 자주 등장하는 문자열을 병합, 이를 어휘 집합에 추가.

원하는 어휘 집합 크기가 될 때 까지 병합.

② 토큰화 : 토큰화 대상 문장의 각 어절에서 어휘 집합에 있는 서브워드가  
포함되었을 때 해당 서브워드를 어절에서 분리.

### - 제주어-한국어 번역을 위한 최적의 BPE 어휘 크기 결정

5개의 어휘 크기(2k, 4k, 8k, 16k, 32k)를 사용 결과 BLEU 점수가 가장  
우수한 **4k** 어휘 크기로 결정.

Lang. Pair	# Vocab.	Dev	Test
kor → jje	2k	44.80	43.26
	<b>4k</b>	<b>44.85</b>	<b>43.31</b>
	8k	44.40	43.03
	16k	43.33	42.08
	32k	42.57	41.07
jje → kor	2k	69.05	67.63
	<b>4k</b>	<b>69.35</b>	<b>67.70</b>
	8k	69.02	67.46
	16k	67.61	66.30
	32k	66.32	65.08

\_그것 은 \_나 를 \_부 유 하게 \_하는 \_좋은 \_효 과 가 \_있습니다 \_.

\_그리고 \_부 자가 \_되 는데 \_어떤 \_좋은 \_점 이 \_있 나요 \_?

\_별 이 \_발 견 되 면 \_추 가 로 \_더 \_많은 \_별 를 \_구 매 할 \_수 \_있 게 \_해 줍 니다 \_.

## PREPROCESS

데이터셋

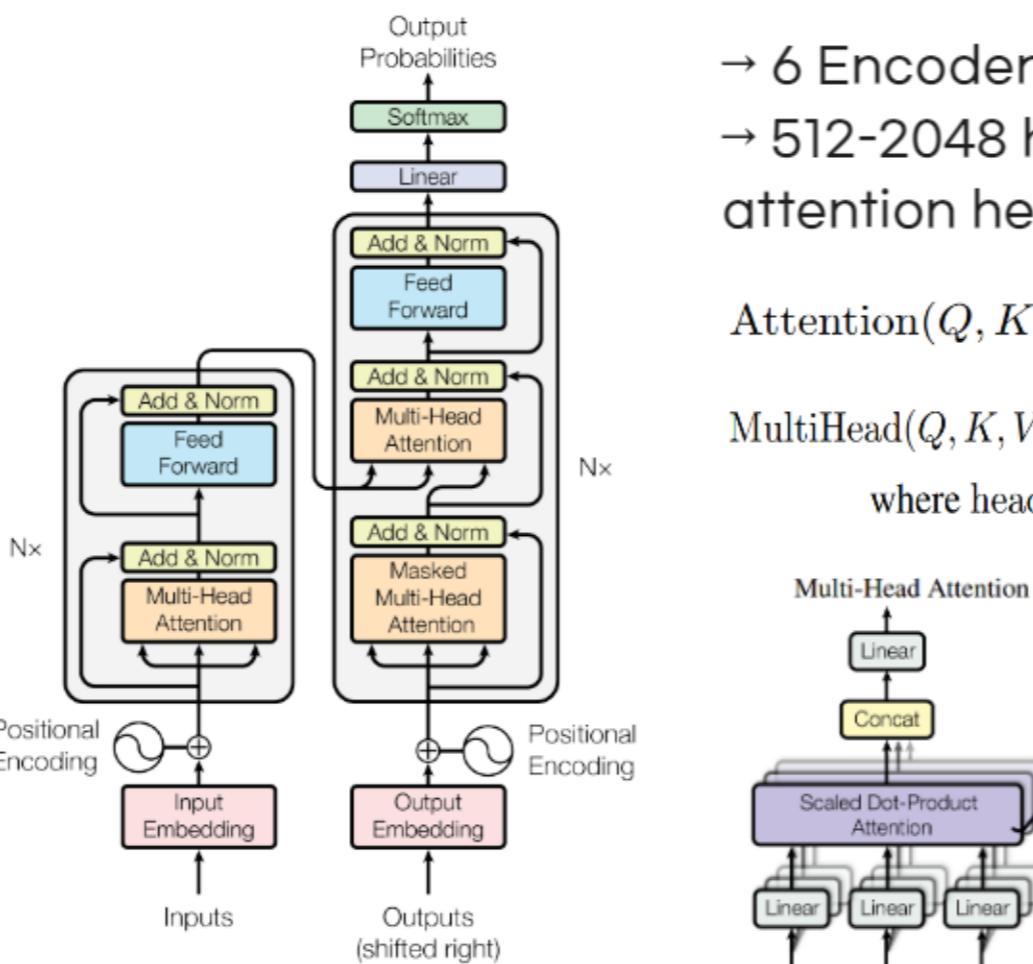


# MODEL

## Transformer

### Sequence-to-Sequecne (one-to-one) based Transformer

Jejueo-Korean Neural Machine Translation



- 6 Encoder and Decoder blocks
- 512-2048 hidden units across 8 attention heads

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$

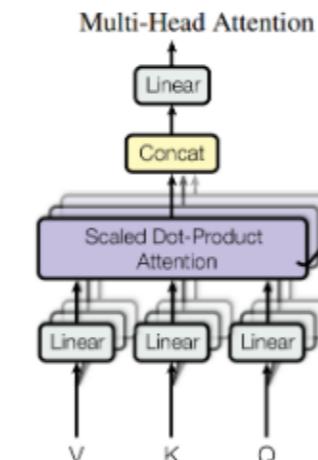


Figure 1: The Transformer - model architecture.

# KOREAN TO JEJU TRANSLATION

*Fairseq*



Translation, Summarization, Language Modeling 및  
기계 번역 Task를 위한 Sequence Modeling Toolkit



# KOREAN TO JEJU TRANSLATION

*Fairseq*

## STEP 01 : BPE Segment (Byte Pair Encoding)

- vocabulary size 4000
- SentencePiece 이용하여 입력 텍스트를 토큰화
- 파라미터 조정을 통해 Unknown token의 경우, 기존 텍스트로 대체

## STEP 02 : fairseq-preprocess

- Build vocabularies and binarize training data

## STEP 03 : fairseq-train

- Train a new model

```
-- optimizer adam  
-- lr 0.0005  
-- dropout 0.3  
-- weight-decay 0.0001  
...
```



# KOREAN TO JEJU TRANSLATION

*Fairseq*

## STEP 04 : fairseq-generate

- Translate pre-processed data with a trained model

S-4084 불은 어떻게 켜는 겁니까 ?

-> 표준어 (input)

T-4084 불은 어떻 싸는 거파 ?

-> 정답 제주어 (target)

H-4084 -014325210452079773 불은 어떻 싸는 거파 ? -> 예측 결과 (output)

D-4084 -014325210452079773 불은 어떻 싸는 거파 ?

P-4084 -0.0479 -0.1122 -0.1495 -0.2253 -0.0990 -0.3056 -0.0997 -0.1068

## STEP 05 : fairseq-score

-> BLEU Score : 44.47



# KOREAN TO JEJU TRANSLATION

***Result***

표준어	제주 사투리
그들은 <b>누구</b> 에게 속해 <b>있습니까?</b>	그들은 <b>누개안티</b> 속해 <b>잇수과?</b>
당신은 그 별들을 소유하고 <b>있습니까?</b>	당신은 그 별덜을 소유 <b>헴수과?</b>
그게 <b>전부</b> 필요한 <b>건가요?</b>	그게 <b>문딱</b> 필요한 <b>거마씨?</b>
<b>어떻게</b> 소통할 거야?	<b>어떻</b> 소통헐 거라?
<b>그리고</b> 그들의 사용자 인터페이스는 <b>왜 그래요?</b>	<b>게고</b> 그들의 사용자 인터페이스는 <b>무사 기꽈?</b>

\* 마씨 : 경어표현으로 ~요 와 비슷하게 쓰임

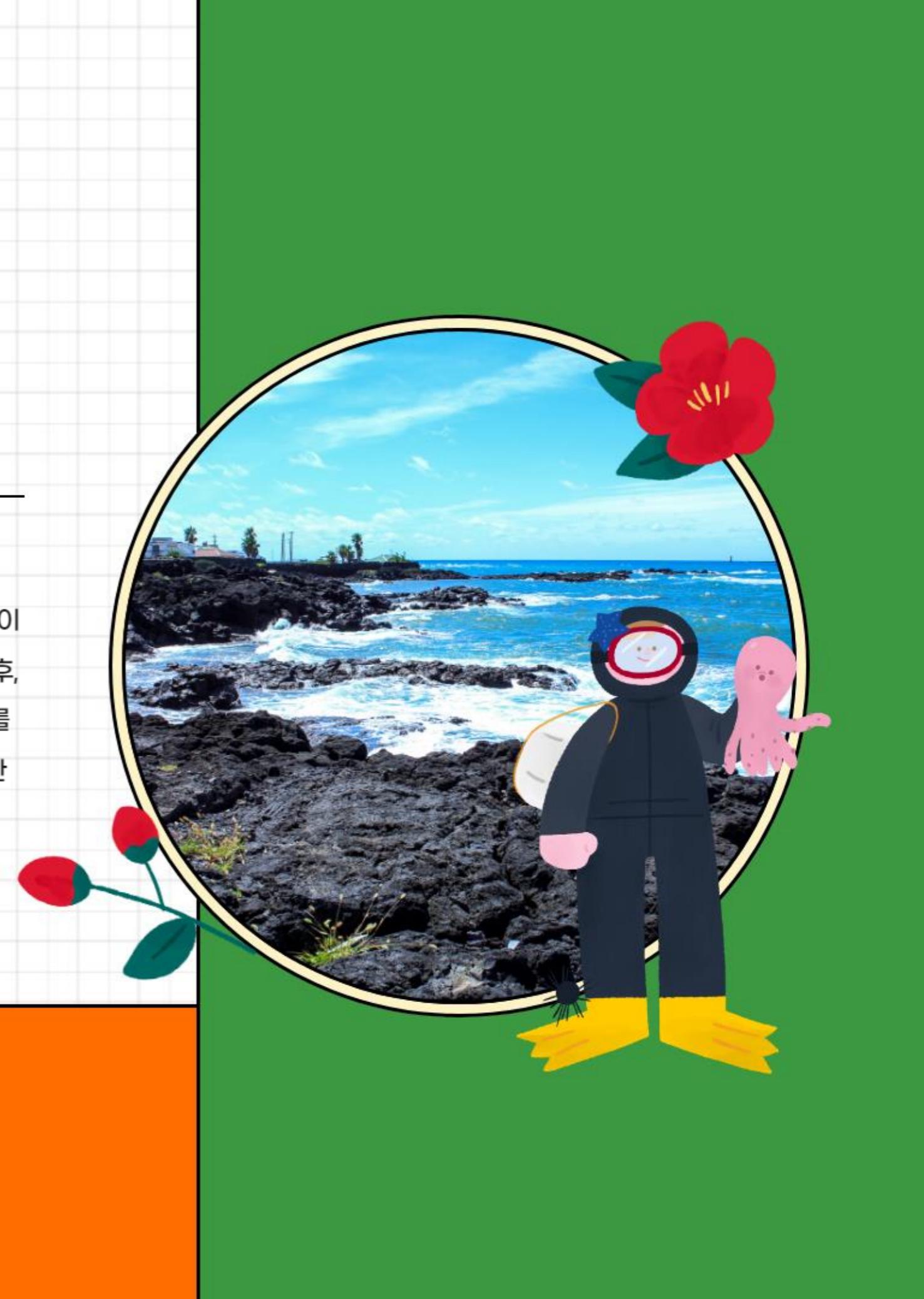
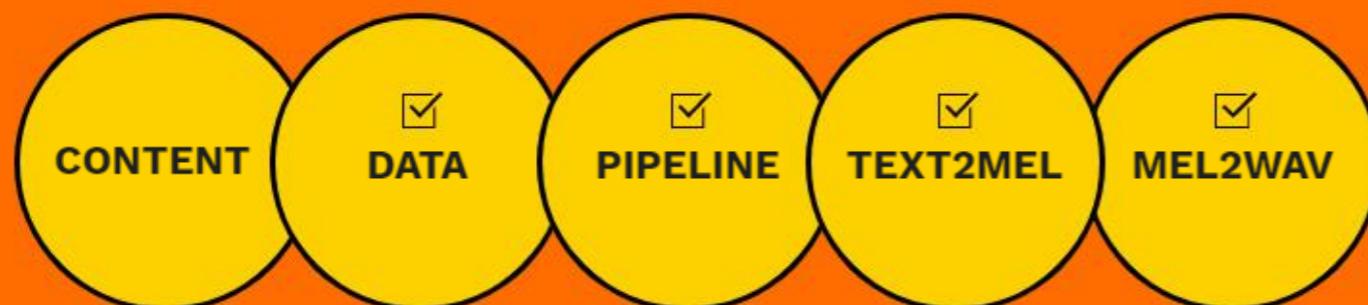


*So, Where we go?*

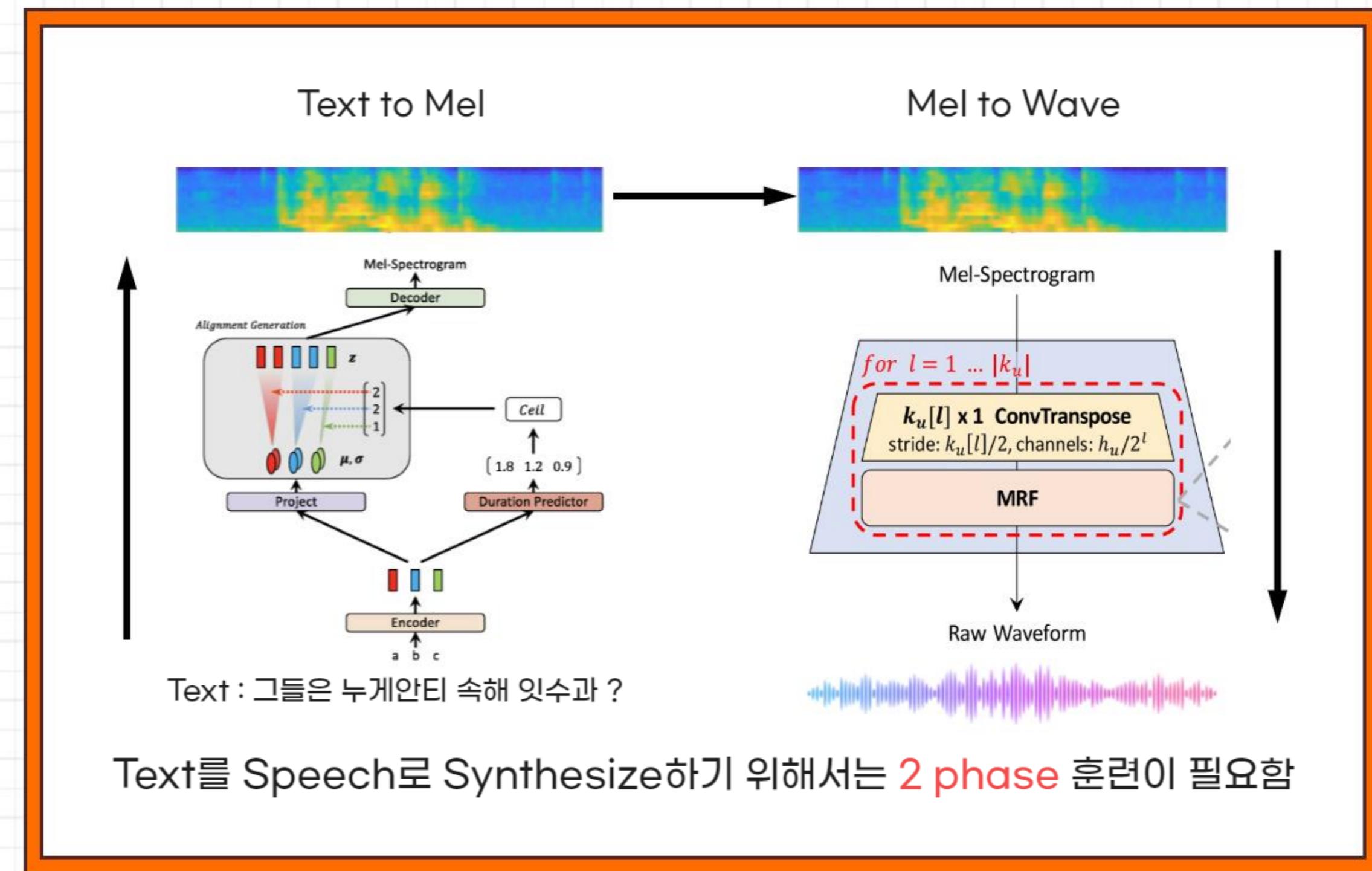
# 제주도 방언 음성합성

**“들리우파? 제주도 방언이?”**

현재 대부분의 Text To Speech (TTS) 모델들은 End-to-End 보다는 2-phase으로 학습하는 것이 성능이 좋다고 알려져 있다. Wav 파일에서 인간의 특징을 반영한 음성 Feature인 Mel-Spectrogram을 추출한 후, Text에서 Mel-Spectrogram을 예측하도록 학습된 모델 (Text to Mel)과 Mel-Spectrogram에서 Wav를 예측하도록 학습된 모델 (Mel to Wav)를 각각 학습해 음성을 합성한다. 원활한 학습을 위해 한 사람이 발화한 제주도 방언 음성 파일과 텍스트 쌍을 활용해 제주도 방언을 학습하는 Text To Speech (TTS) Pipeline을 학습해보자.



# TEXT TO SPEECH PIPELINE



## Kakao 제주도 방언 오디오 데이터

Kakao Jejueo Datasets for Machine  
Translation and Speech Synthesis

카카오 브레인은 제주도 방언 인터뷰 스크립트를  
통해 만든 JIT(Jejuo Interview Transcripts)  
중 10,000개를 제주도 방언 오디오 데이터 세트인  
JSS(Jejueo Single Speaker Speech) 개발.

kakaobrain

보다 깨끗한 소리를 합성하기 위해 Wav File을 Trimming 하고,  
기존의 Sampling Rate를 22050으로 변경한 후에 Mel-Spectrogram을 추출



jss/0.wav	지금부터 그러면 본격적으로 우리가 도련동에 대해서 조사를 할 거라 예? 이 마을이 ...
jss/1.wav	예, 그건 한 칠백년 전에 이제 그 설촌이 시작이 되었다고 헝니다.
jss/2.wav	예. 칠백년 전에 설촌이 뛰었는데 이제 그루후에 이제 성씨들이 여러 성씨들이 많이 ...
jss/3.wav	예. 그러면은 칠백년부터 하는데 설촌할 때 어떤 성씨들이 했던 말도 이신가마씨?
jss/4.wav	그 다음 양씨. 고씨. 마 대략적으로 요런 순서가 됐서양.
jss/5.wav	예. 게은 그때는 그렇게 하다가 지금은 어느 성씨가 절 하우파? 요즘 은, 요즘 도련?
jss/6.wav	요즘은 역시 김씨, 고씨 이제 그렇게덜 잊어난디 이제 그 사방으로 이 젊은 사름이 ...
jss/7.wav	이젠 성씨가 아까 뭐 몇 가지지?
jss/8.wav	그것도 그 성씨가 내가 그전에 반장을 헤봤는데.
jss/9.wav	반장을 여기서 여기 도련에 일반인디 여기가.
jss/10.wav	우리 반 귀역만 조사해 보니까 한 이십여 성씨가 되니까, 아 이젠 그것에 충미가 부...
jss/11.wav	상당이 많은 성씨 가짓는데 기씨같은 분들은 에 대개 그 여자 어른 멀이 앗다가 이제...
jss/12.wav	그 지금 그러면 지금 여기가 도련일동 아니우까예? 여기가 옛날 이름 뭐라마씨?
jss/13.wav	지금 도련드르는 게면 몇 명 정도 살고 있고 아까 일반이라고 했는데 몇 개 반으로 ...

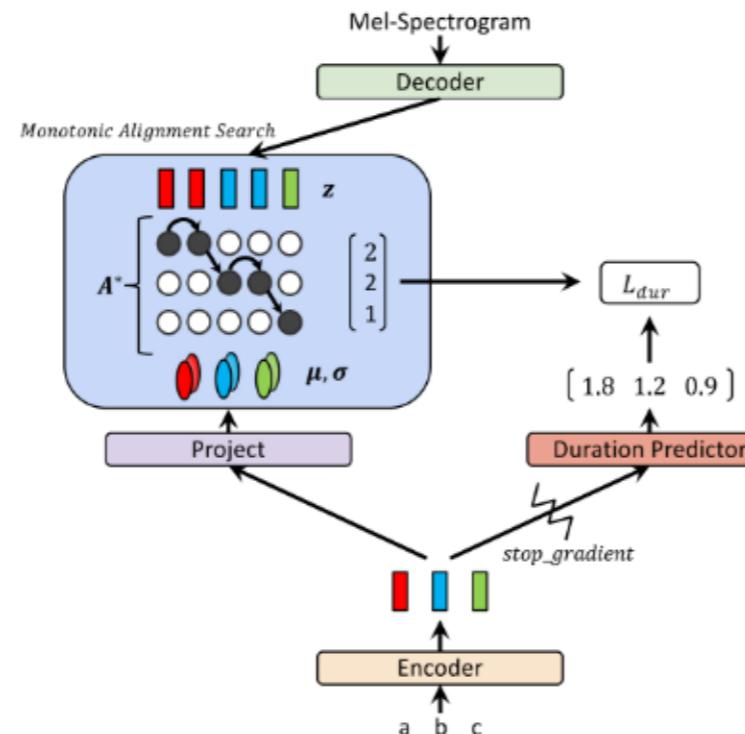
## DATASET — 데이터셋



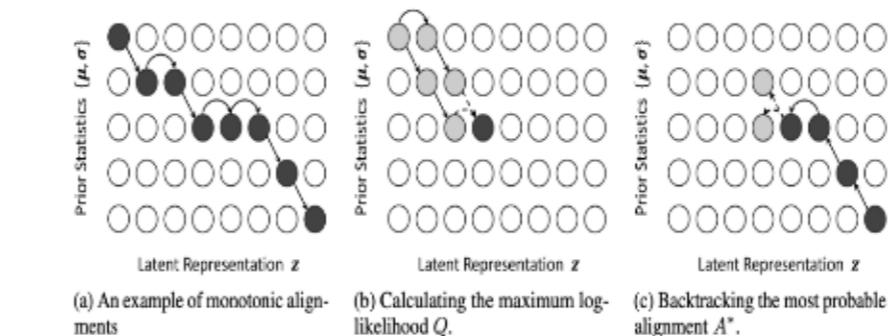
# TEXT TO MEL SPECTROGRAM

**Glow TTS**

## Structure of Glow TTS



## Monotonic Alignment Search



## Objective Function of Glow TTS

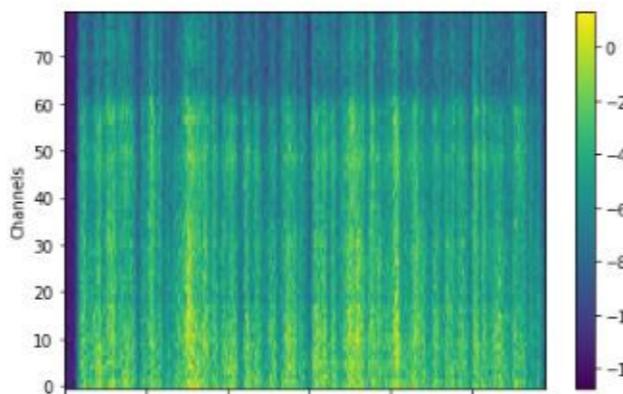
$$\max_{\theta, A} L(\theta, A) = \max_{\theta, A} \log P_X(x|c; A, \theta)$$
$$A^* = \arg \max_A \log P_X(x|c; A, \theta) = \arg \max_A \sum_{j=1}^{T_{mel}} \log \mathcal{N}(z_j; \mu_{A(j)}, \sigma_{A(j)})$$

Tacotron2와 같은 Auto-Regressive한 모델의 한계를 극복하기 위해 개발된 **Flow-based Model**로,  
Dynamic Programming 기반의 Alignment Search를 통해 MLE관점에서 Monotonicity와 Surjection을  
만족시키는 Text & Mel-Spectogram 간의 Alignment( $=A^*$ )을 찾은 후 Gradient Descent를 통해 최적화하는 모델

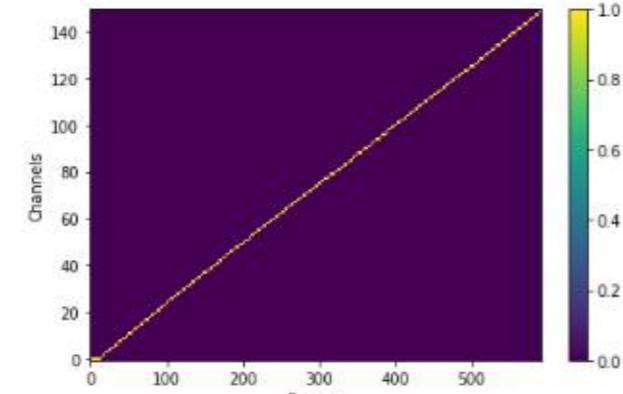
# TEXT TO MEL SPECTROGRAM

***Result***

Text : 지금부터 그러면 본격적으로 우리가 도련동에 대해서 조사를 할 거라 예?  
이 마을이 언제 어떻게 형성됐던 헌 말 알아지는 데로 ㄱ ㅋ 아줍서.

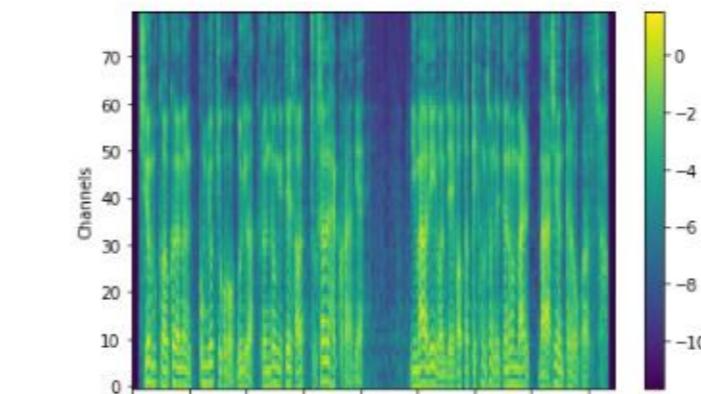


Mel-Spectrogram

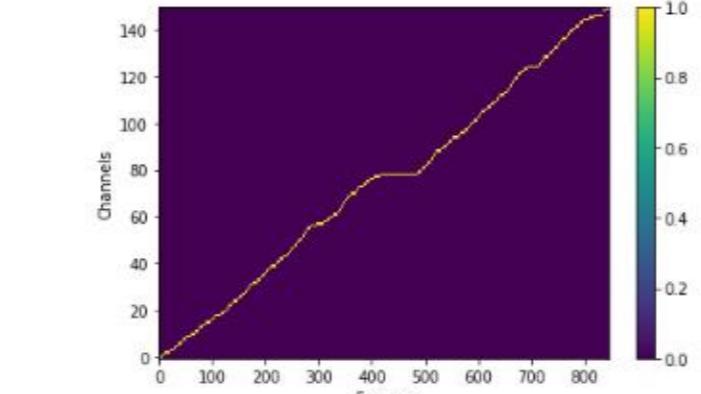


Attention Alignment

Epoch 01



Mel-Spectrogram



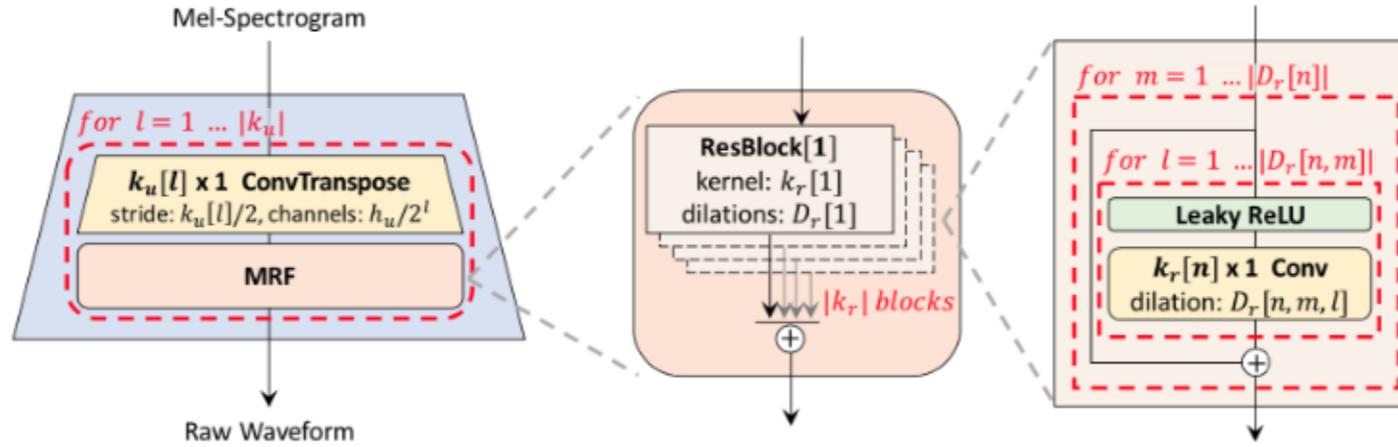
Attention Alignment

Epoch 40

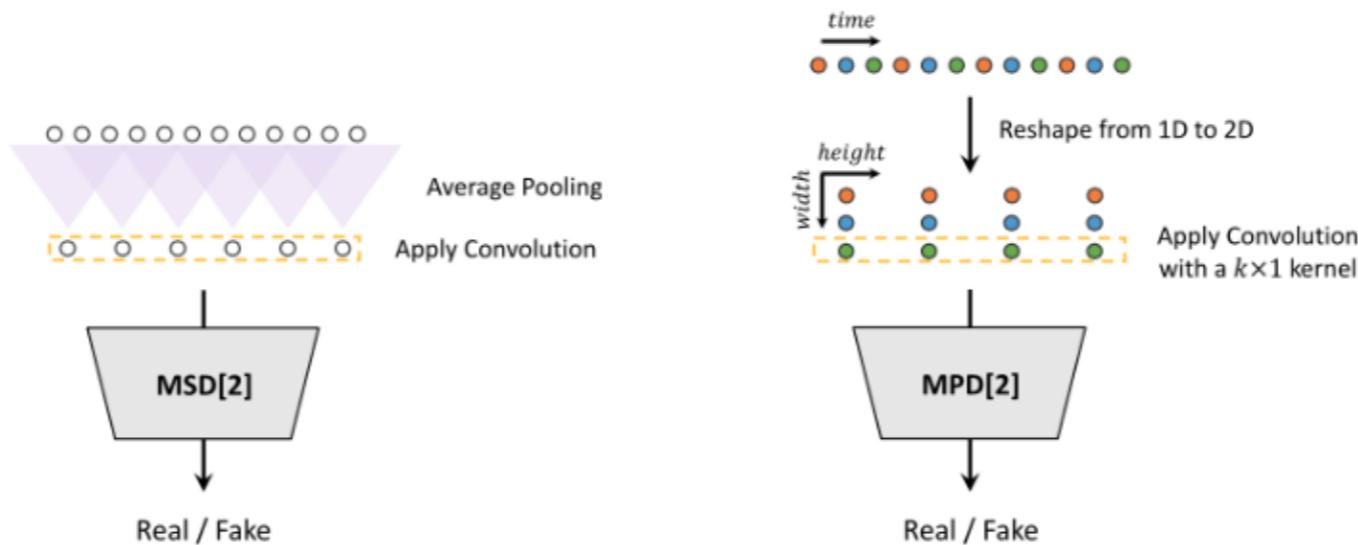
# MEL SPECTROGRAM TO WAVEFORM

**hifi-GAN**

기존의 AR모델보다 학습 소요 시간이 짧고, quality도 높은 모델!



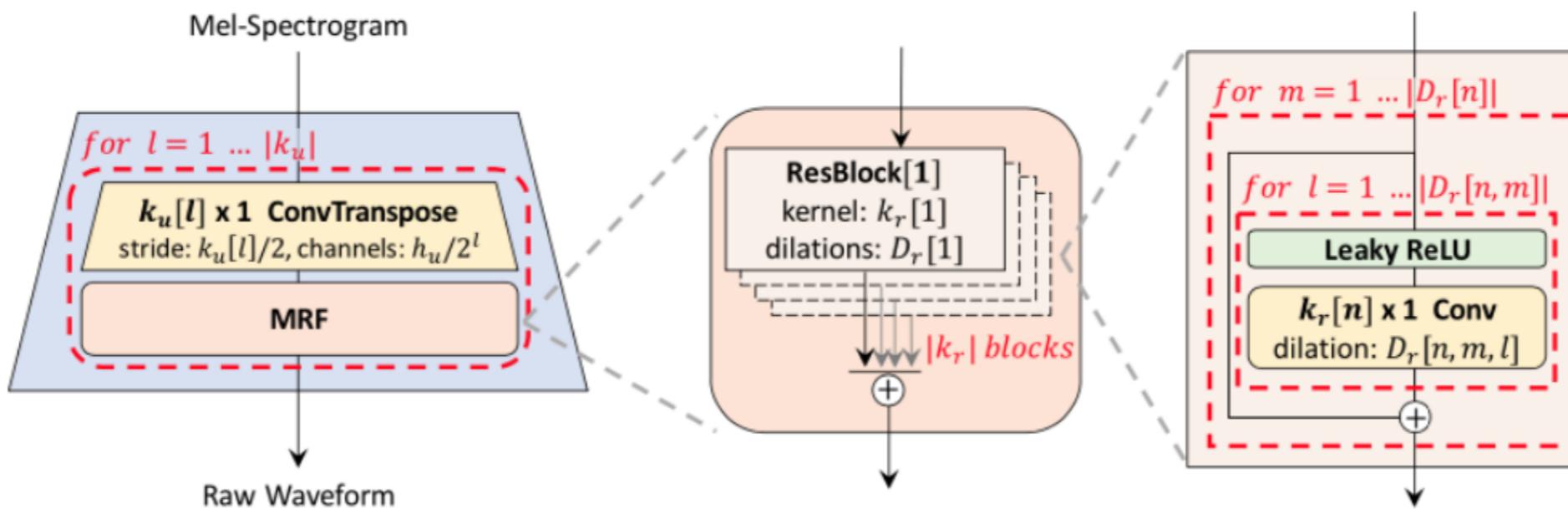
1개의 generator와 2개의 discriminator로 구성



# MEL SPECTROGRAM TO WAVEFORM

**hifi-GAN**

Generator



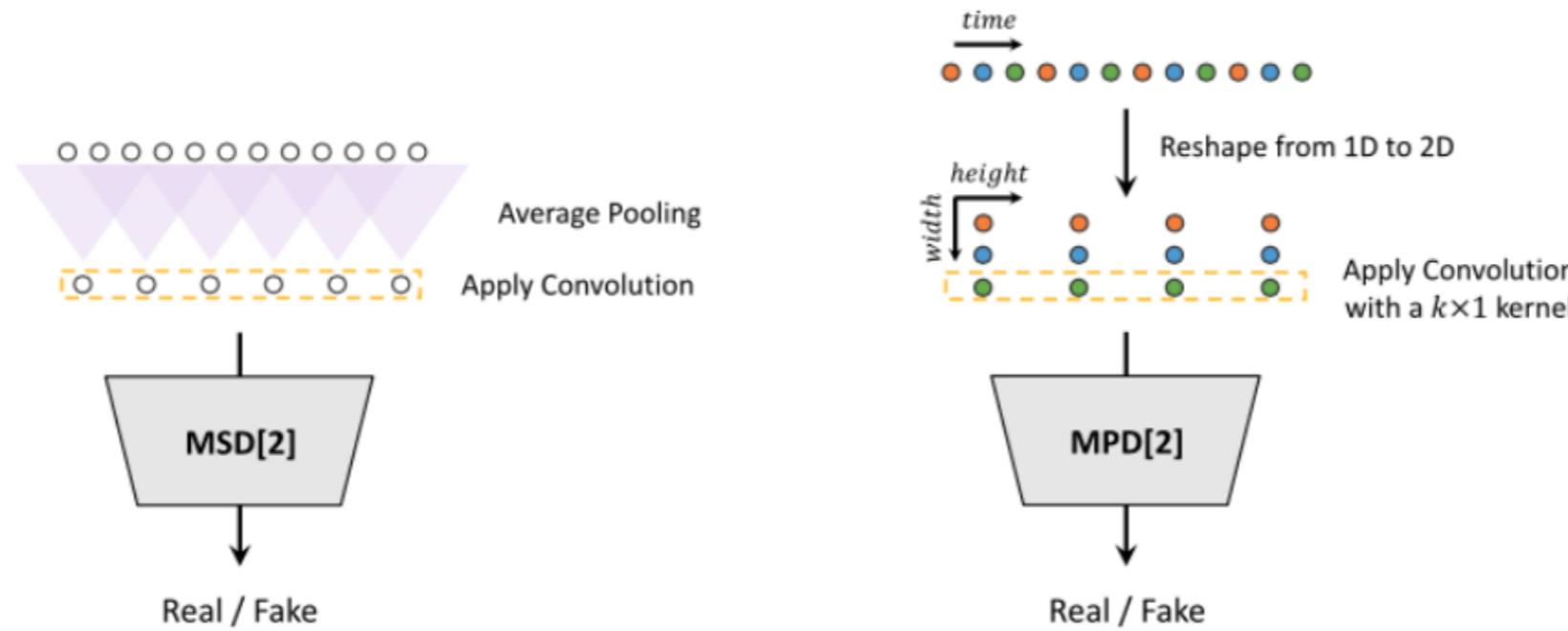
Step1 : Mel-spectrogram을 입력으로 받아 upsampling

Step2 : MRF모듈을 거쳐 병렬적으로 다양한 길이의 패턴을 관찰

# MEL SPECTROGRAM TO WAVEFORM

***hifi-GAN***

Discriminator



MPD : 특정 주기로 일정하게 나뉜 입력 음성을 다루는 sub-discriminator로 구성

T길이의 1D raw audio를 높이 T/p, 너비 p의 2D데이터로 변환

p각 sub-discriminator는 입력 음성의 다른 부분을 받아 implicit structure를 찾음

MSD : 연속된 음성을 평가하기 위해 input scale를 조정한 sub-discriminator로 이루어짐

input scale을 조정해 1, 1/2, 1/4에서 동작하는 sub-discriminator 3개로 구성

# MEL SPECTROGRAM TO WAVEFORM

***hifi-GAN***

Objective Function

## GAN Loss

Non-vanishing gradient flow를 위해 binary cross-entropy를 Least square loss function으로!

$$\mathcal{L}_{Adv}(D; G) = \mathbb{E}_{(x,s)} \left[ (D(x) - 1)^2 + (D(G(s)))^2 \right] \quad (1)$$

$$\mathcal{L}_{Adv}(G; D) = \mathbb{E}_s \left[ (D(G(s)) - 1)^2 \right] \quad (2)$$

## Mel-Spectrogram Loss

Generator의 학습 효율을 높이고, 생성된 음성의 품질을 높이기 위해 추가한 loss

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(x,s)} \left[ \|\phi(x) - \phi(G(s))\|_1 \right] \quad (3)$$

## Feature Matching Loss

GT sample과 generated sample 사이의 feature of discriminator의 차이로 측정

$$\mathcal{L}_{FM}(G; D) = \mathbb{E}_{(x,s)} \left[ \sum_{i=1}^T \frac{1}{N_i} \|D^i(x) - D^i(G(s))\|_1 \right] \quad (4)$$

# MEL SPECTROGRAM TO WAVEFORM

***hifi-GAN***

Objective Function

Objective function Minimize G

GAN Loss + Feature Matching Loss + Mel-Spectrogram Loss

$$\mathcal{L}_G = \mathcal{L}_{Adv}(G; D) + \lambda_{fm}\mathcal{L}_{FM}(G; D) + \lambda_{mel}\mathcal{L}_{Mel}(G) \quad (5)$$

Objective functino Maximize D

GAN Loss

$$\mathcal{L}_D = \mathcal{L}_{Adv}(D; G) \quad (6)$$

Considering even the sub-discriminator of D

5번 식에서 sub discriminator까지 고려

$$\mathcal{L}_G = \sum_{k=1}^K \left[ \mathcal{L}_{Adv}(G; D_k) + \lambda_{fm}\mathcal{L}_{FM}(G; D_k) \right] + \lambda_{mel}\mathcal{L}_{Mel}(G) \quad (7)$$

$$\mathcal{L}_D = \sum_{k=1}^K \mathcal{L}_{Adv}(D_k; G) \quad (8)$$

# MEL SPECTROGRAM TO WAVEFORM

***hifi-GAN***

training details

hyperparameter of hifiGAN

```
--Batch size 16
--Learning rate 0.0002
-- Upsample_rates [8,8,2,2]
--Upsample_kernel_sizes [16,16,4,4]
--Upsample_initial_channel 512
--Resblock_kernel_size [3,7,1]
--Resblock_dilation_size [[1,3,5],[1,3,5],[1,3,5]]
--Num_mel 80
--Sampling_rate 22050
--Training step 200000
```

Training details



```
--GPU P100(24GB)
--Training time 10days
--loss_gan_all 28
--mel_error 0.29
--pretrained_model
False
```

# FINAL OUTPUT OF JEJUDO-TTS

**English**

"Kings do not own, they reign over. It is a very different matter."

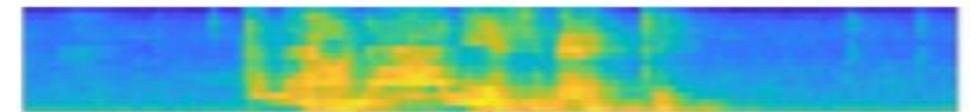
**Korean**

"왕은 소유하지도 않고 지배를 해요. 아주 다른 문제죠."

**Jeju Dialect**

왕은 소유허지도 안허고 지베를 헤여 . 고 아주 다른 문제라예 .

**Mel Spectrogram**



**Speech**



# OUR FANTASTIC OUTPUT



# STEVE JOBS



# LITTLE PRINCE

어린 왕자

The Little Prince

# CONCLUSION

## Limitation

- Pre-trained된 mBART를 사용하면 빈도수가 낮은 고유명사는 정확히 번역하지 못하는 어려움이 있음
- 제주도 음성 데이터의 경우 데이터 자체의 품질이 좋지 않아 선명한 음성을 합성하기에 한계가 있음
- 사용할 수 있는 자원이 제한적임 (Colab pro 기준, P100 1대)
- 시간의 부족으로 hifi-GAN의 경우 200,000 step까지 밖에 학습하지 못함

## Further work

- Eng-Kor & Kor-Jeju Translation는 copy-mechanism을 추가해 빈도수가 낮은 고유명사는 원문의 표현을 그대로 활용하도록 하는 작업이 필요함
- hifi-GAN의 경우 일반적으로 선명한 음성 합성을 위해 요구되는 훈련시간 (평균 250만 step)까지 지속적인 학습 필요

# REFERENCES

- Park, K., Choe, Y. J., & Ham, J. (2019). Jejueo Datasets for Machine Translation and Speech Synthesis. arXiv preprint arXiv:1911.12071.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742.
- Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. Advances in Neural Information Processing Systems, 33, 8067-8077.
- Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33, 17022-17033.
- Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038.

# WE ARE ...





[표준어] 안녕히 계세요 여러분.  
[제주어] 펜안이 이십서 여러분.