고려대학교
빅데이터 연구회

KU-BIG

너의 목소리가 보여

딥러닝 2조

이찬희 박인성 이강현 이민수 정희정

INDEX

1. TTS : What's the use

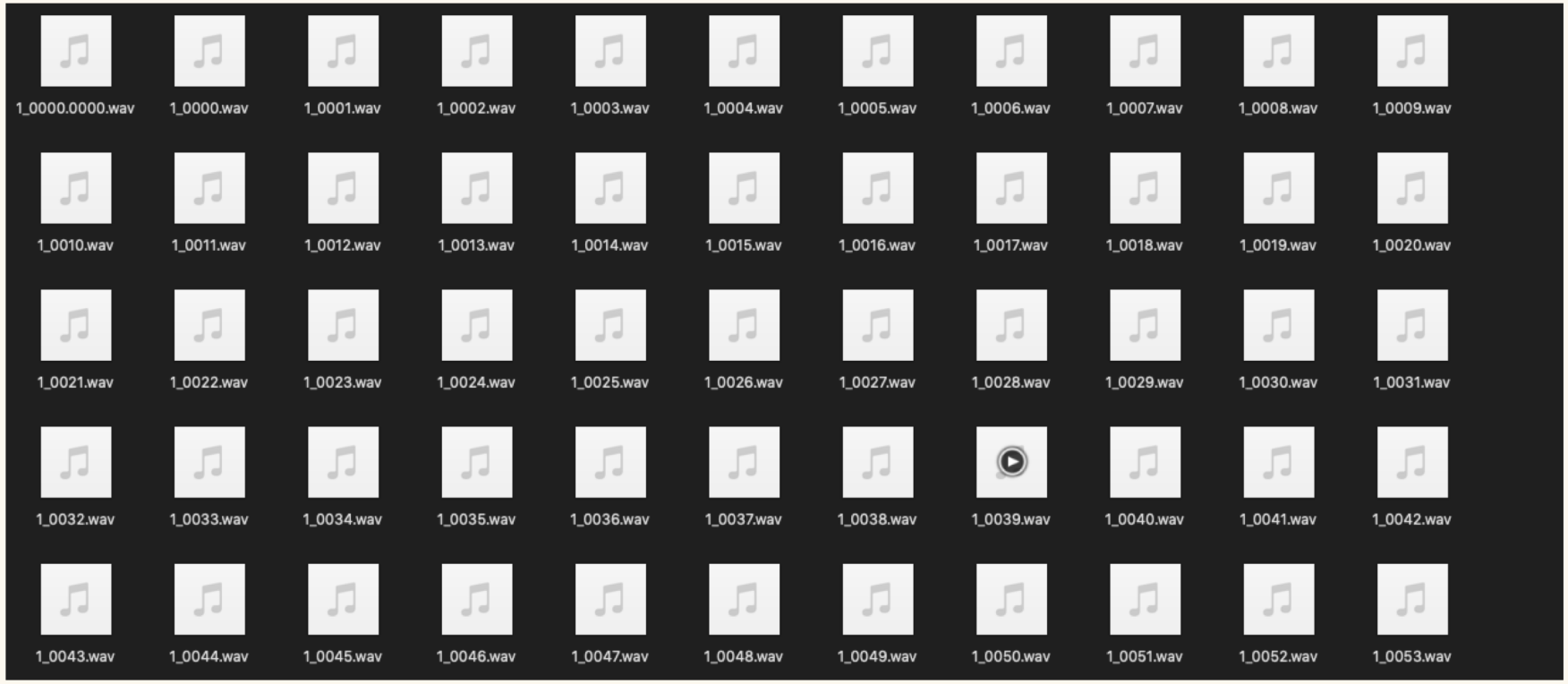
카카오, 네이버의 AI 스피커의 출시 및 상용화

영화 속 TTS 기술

2. Korean Single Speaker

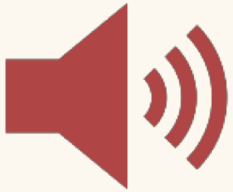
약 13시간 총 12,853개 음성 DATA

2. Korean Single Speaker



개별 6초 - 15초 사이의 음성 파일 + Script

3. 데이터 전처리



Librosa

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

1개의 sample = 152*20 행렬

sampling Rate : 22500

1 초의 음성 → 22500*152*20 의 3차원 배열

t 초의 음성 → 22500t*152*20 의 3차원 배열

3. 데이터 전처리

```python
num_to_kor = {
        '0': '영',
        '1': '일',
        '2': '이',
        '3': '삼',
        '4': '사',
        '5': '오',
        '6': '육',
        '7': '칠',
        '8': '팔',
        '9': '구',
}

unit_to_kor1 = {
        '%': '퍼센트',
        'cm': '센치미터',
        'mm': '밀리미터',
        'km': '킬로미터',
        'kg': '킬로그람',
}
upper_to_kor = {
        'A': '에이',
        'B': '비',
```

```python
etc_dictionary = {
        '2 30대': '이삼십대',
        '20~30대': '이삼십대',
        '20, 30대': '이십대 삼십대',
        '1+1': '원플러스원',
        '3에서 6개월인': '3개월에서 육개월인'
```

"만으로 36 살입니다.",

"A: 몇 년생이세요? B: 78년생이에요.",

"그 사람은 30대 초반이에요.",

"범인은 20대 중반에서 후반의 남성으로 보입니다.",

"저는 평생을 독신으로 살 생각입니다.",

```python
count_checker =
"(시|명|가지|살|마리|포기|송이|
```

```python
count_to_kor1 = [""] +
    ["한","두","세","네","다섯","여섯","
    곱","여덟","아홉"]

count_tenth_dict = {
        "십": "열",
        "두십": "스물",
        "세십": "서른",
        "네십": "마흔",
        "다섯십": "쉰",
        "여섯십": "예순",
        "일곱십": "일흔",
```

```python
kubig_dictionary = {
'KuBig': '쿠빅',
'Bigdata': '빅데이터',
'Deep learning': '딥러닝',
```

심화스터디 내용

1. PreNet 에서 이용되는 CNN개념 학습

: stride, pooling 등의 내용 학습

2. Encoder, Decoder에서 이용되는 RNN개념 학습

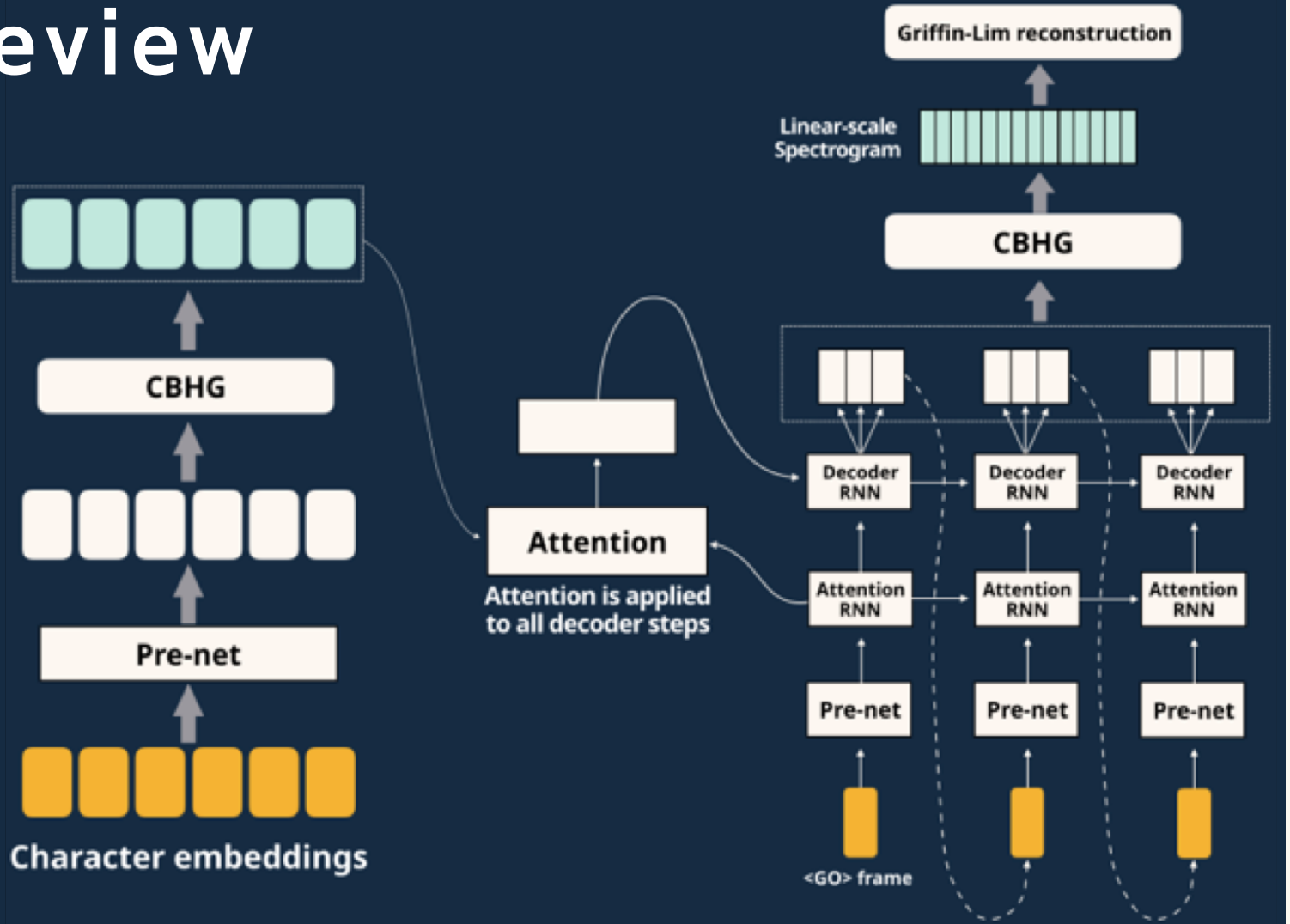
: Toy example로 RNN, LSTM, Attention 구현

3. 코드학습

: Architecture 코드를 이용한 Python 코드 학습

Tacotron : review

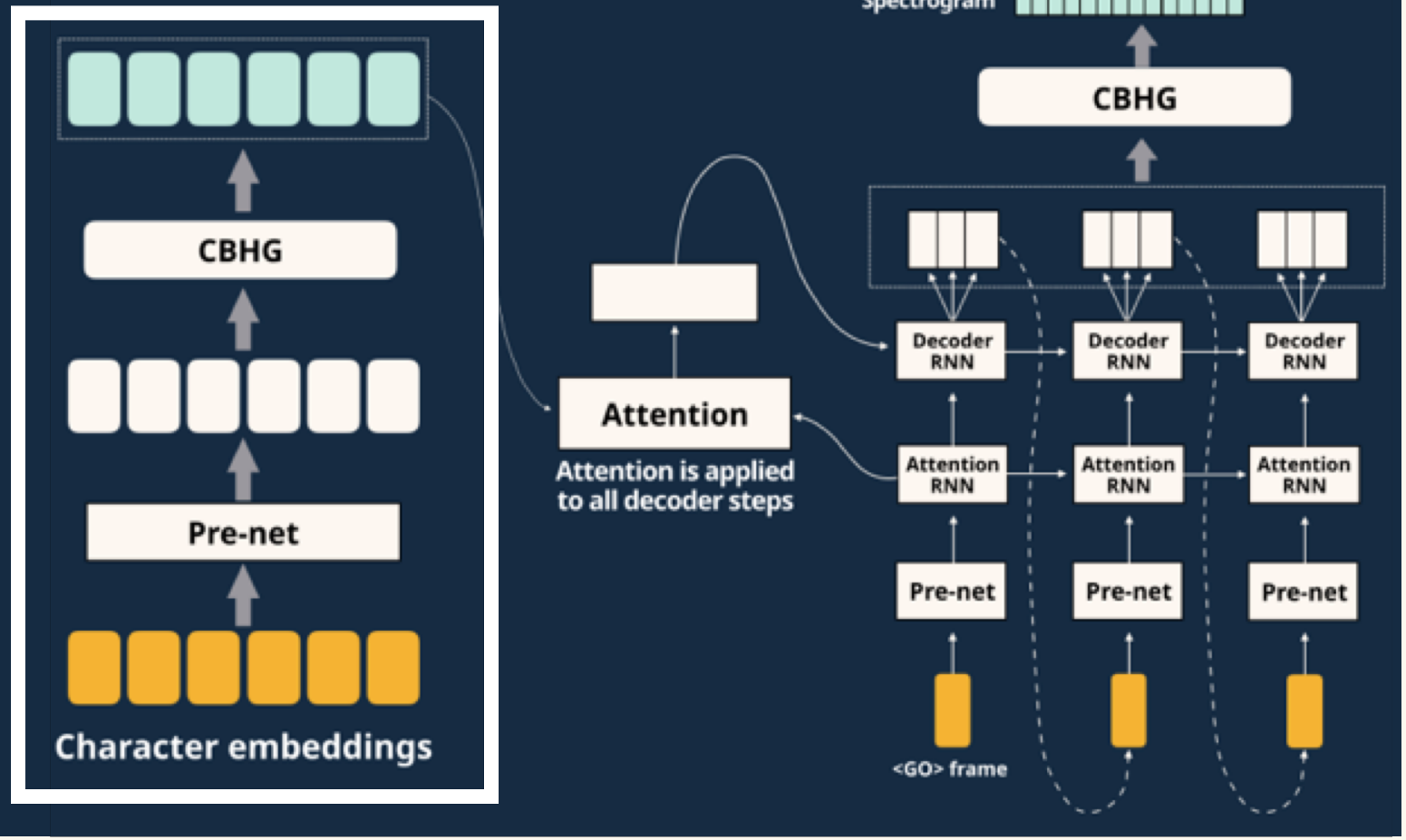
1. ENCODER
2. DECODER
3. ATTETION
4. VOCODER



Tacotron : review

1. ENCODER
2. DECODER
3. ATTETION
4. VOCODER

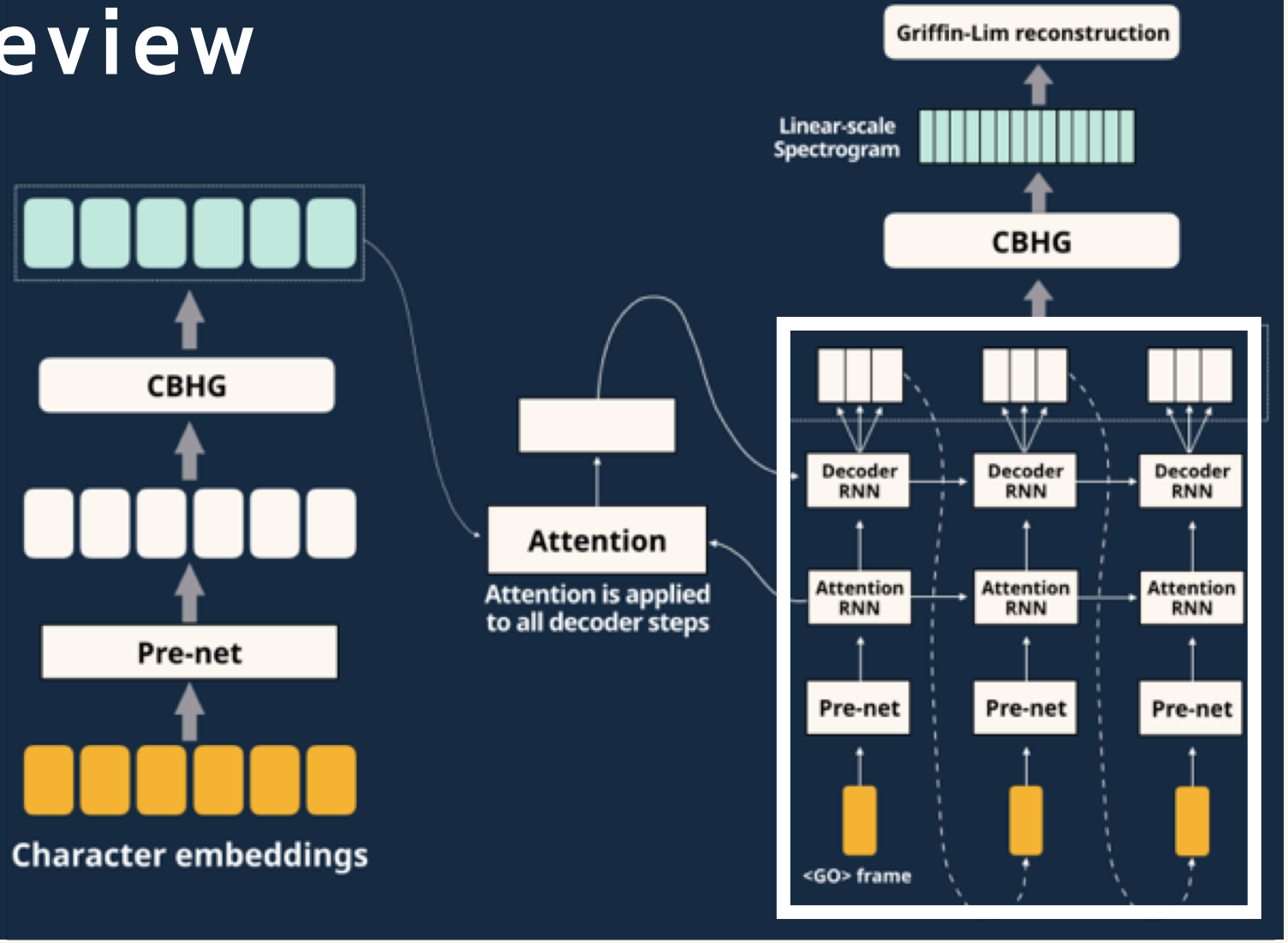
학습해야하는 음성을 기계가 잘 알아들을 수 있도록 숫자로 바꾸고,



Tacotron : review

1. ENCODER
2. DECODER
3. ATTETION
4. VOCODER

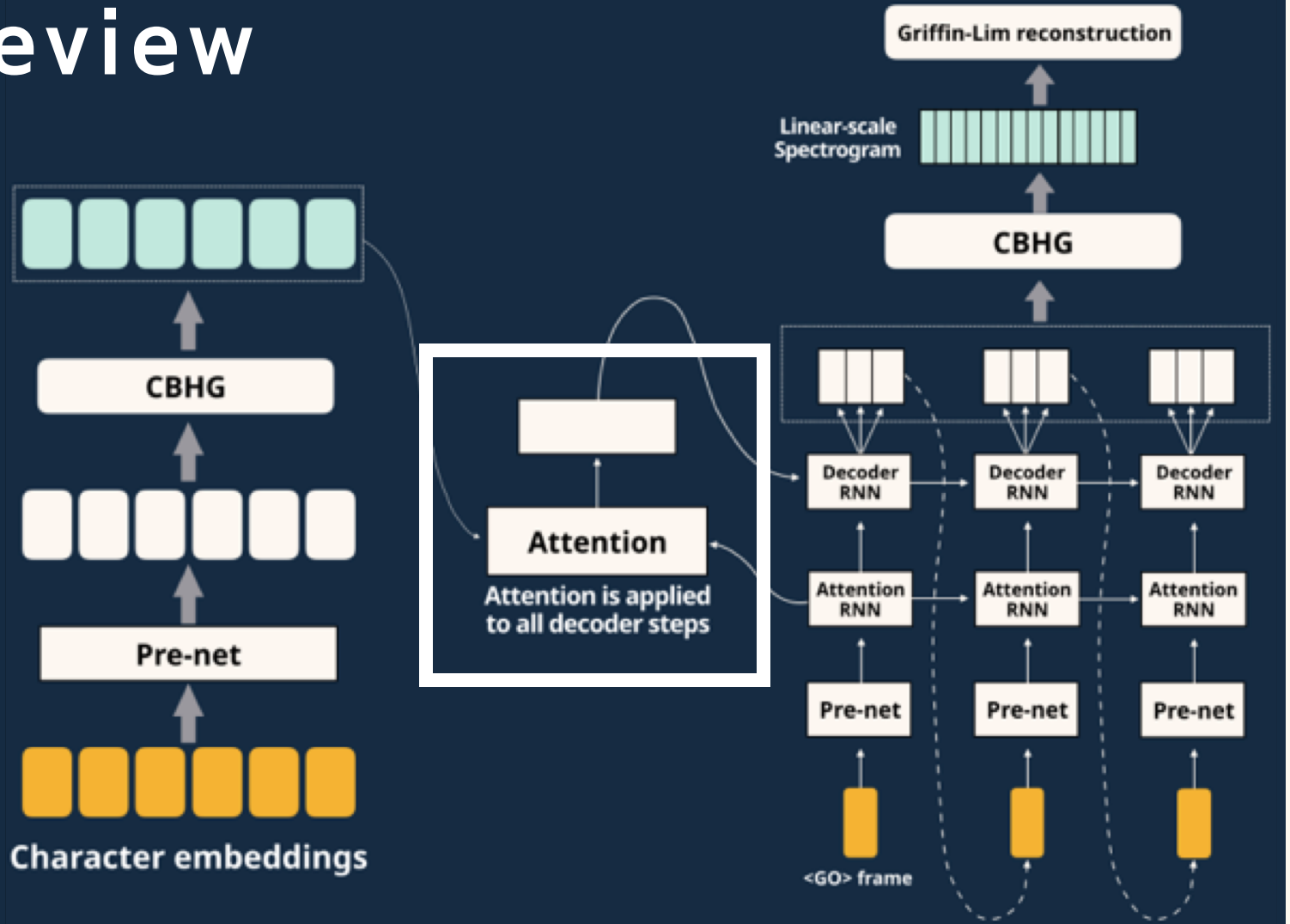
이 숫자를
음성으로 구현하기 위해
스펙토그램을 만들고



Tacotron : review

1. ENCODER
2. DECODER
3. ATTETION
4. VOCODER

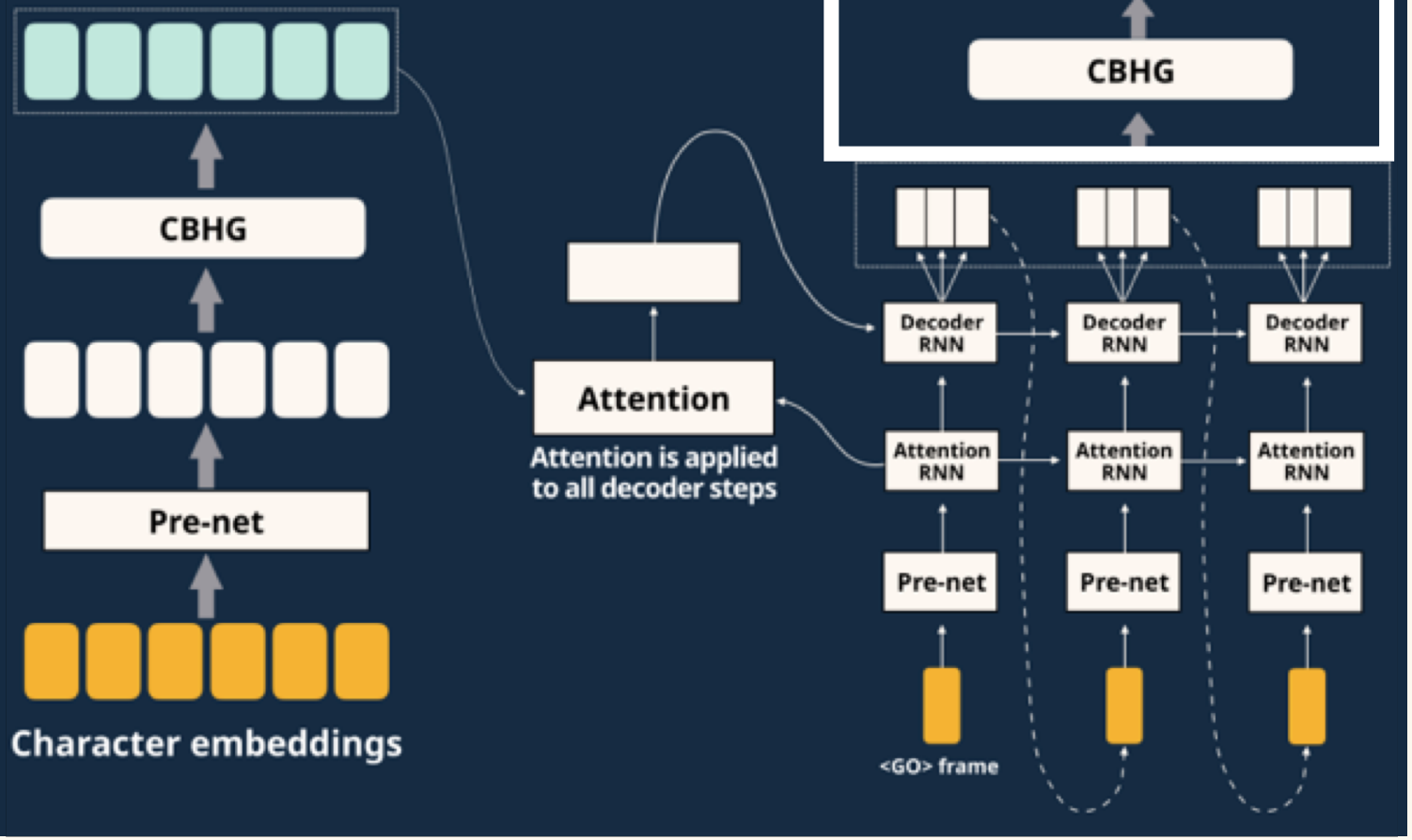
Encoder와
Decoder사이에서
Attention이
적절하게 학습하면서



Tacotron : review

1. ENCODER
2. DECODER
3. ATTETION
4. VOCODER

스펙토그램을
음성으로 만들어준다!



4. Modeling and Performance

Batch_size: 32

Dec_layer: 2

Dropout_prob: 0.5

Enc_stride: 2

Sample_rate: 22050

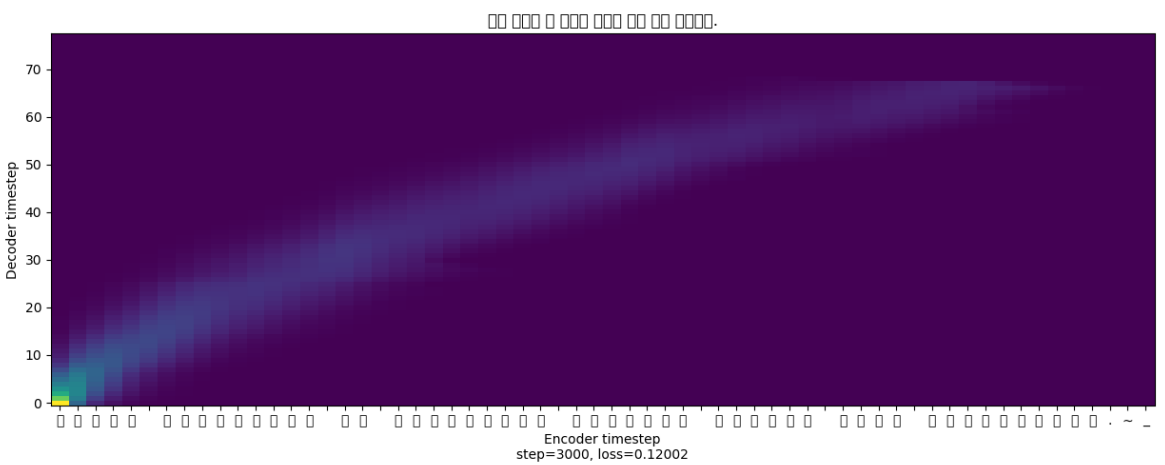
Optimizer: Adam

```
Step 1      [53.587 sec/step, loss=0.12831, avg_loss=0.12831]
Step 2      [34.827 sec/step, loss=0.12984, avg_loss=0.12907]
Step 3      [29.615 sec/step, loss=0.12951, avg_loss=0.12922]
Step 4      [25.698 sec/step, loss=0.12971, avg_loss=0.12934]
Step 5      [25.047 sec/step, loss=0.12722, avg_loss=0.12892]
Step 6      [23.733 sec/step, loss=0.12888, avg_loss=0.12891]
Step 7      [22.456 sec/step, loss=0.12902, avg_loss=0.12893]
Step 8      [21.418 sec/step, loss=0.13006, avg_loss=0.12907]
Step 9      [20.615 sec/step, loss=0.13106, avg_loss=0.12929]
```

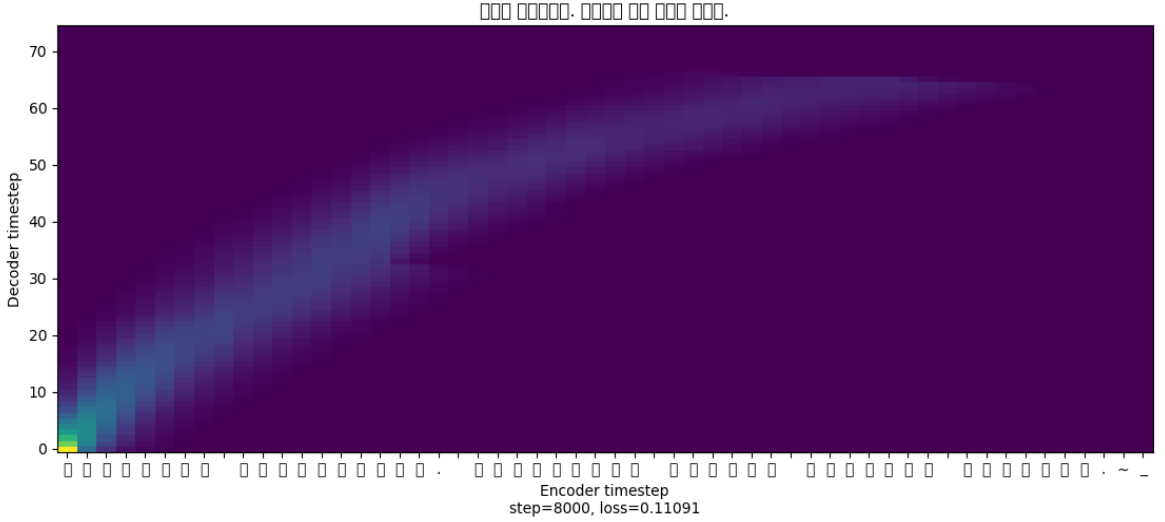
```
Step 5492   [14.632 sec/step, loss=0.10335, avg_loss=0.09871]
Step 5493   [14.606 sec/step, loss=0.10014, avg_loss=0.09870]
Step 5494   [14.617 sec/step, loss=0.09873, avg_loss=0.09870]
Step 5495   [14.630 sec/step, loss=0.09853, avg_loss=0.09869]
Step 5496   [14.692 sec/step, loss=0.09816, avg_loss=0.09867]
Step 5497   [14.664 sec/step, loss=0.09806, avg_loss=0.09867]
Step 5498   [14.686 sec/step, loss=0.09784, avg_loss=0.09865]
Step 5499   [14.592 sec/step, loss=0.10003, avg_loss=0.09866]
Step 5500   [14.564 sec/step, loss=0.10045, avg_loss=0.09868]
```

4. Modeling and Performance

Epoch : 15000

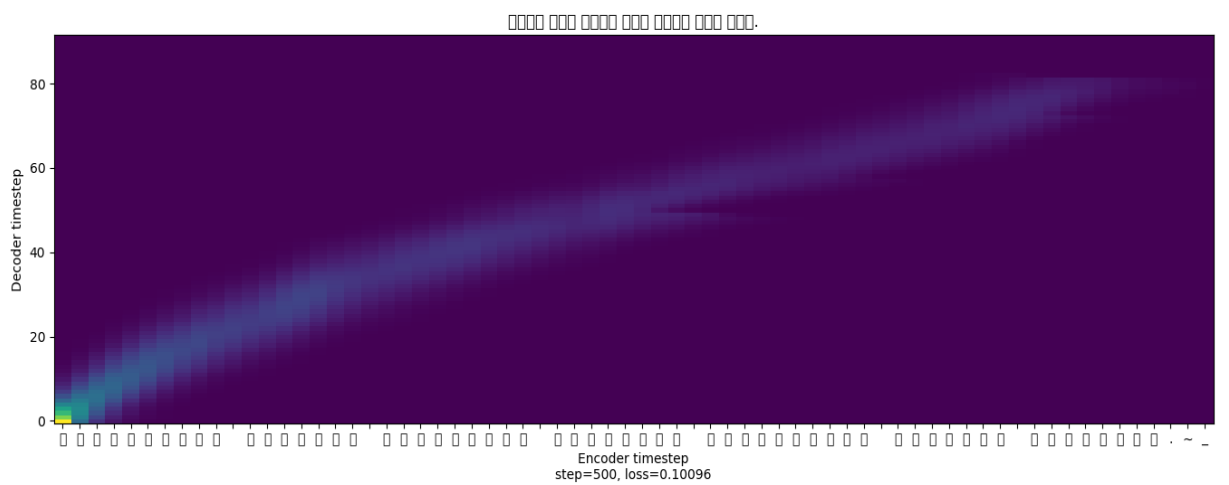


Epoch : 20000

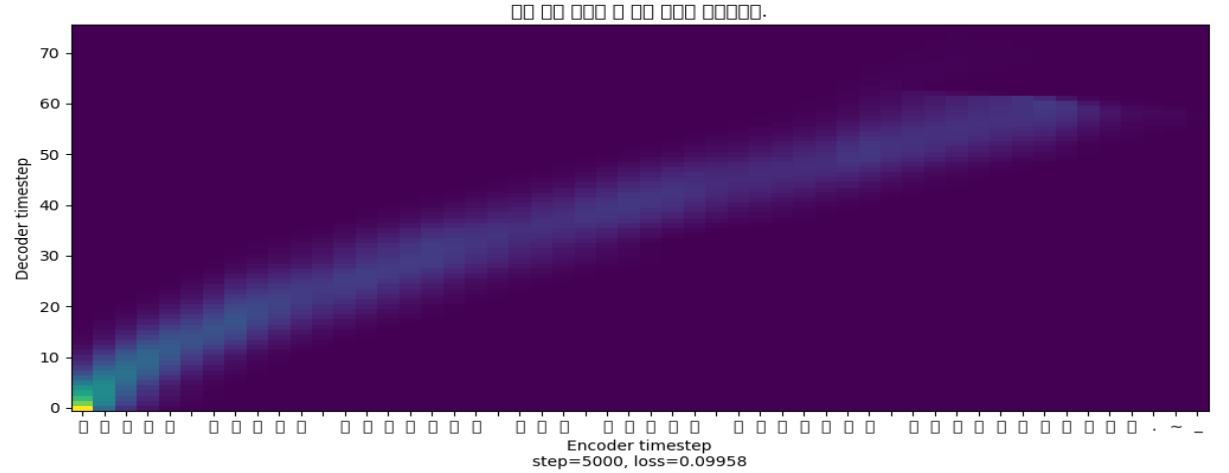


4. Modeling and Performance

Epoch : 25000



Epoch : 30000



Training / Testing Audio

"제가 중학생때 ???? ??? ??."

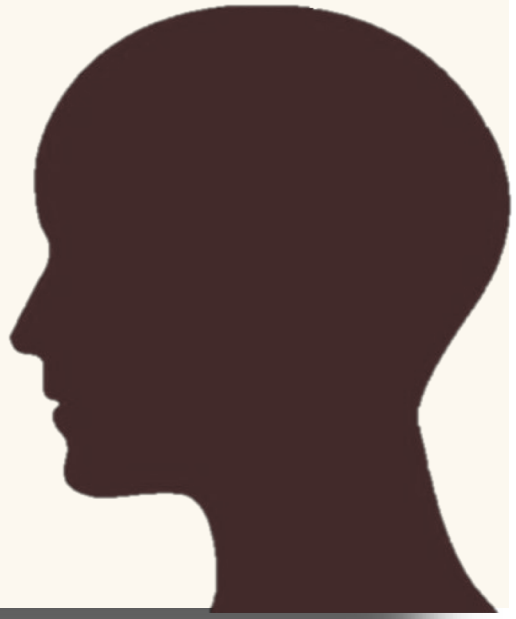
<train audio at 25,000 Epoch>

"잠시만 기다리세요. 주전자에 물을 끓이고 있어요"

<train audio at 30,000 Epoch>

"의사들은 누나를 살리려고 최선을 다했지만,
소용이 없었다."

<train audio at 35,000 Epoch>



Training / Testing Audio

“????? ???? ??? ??.”

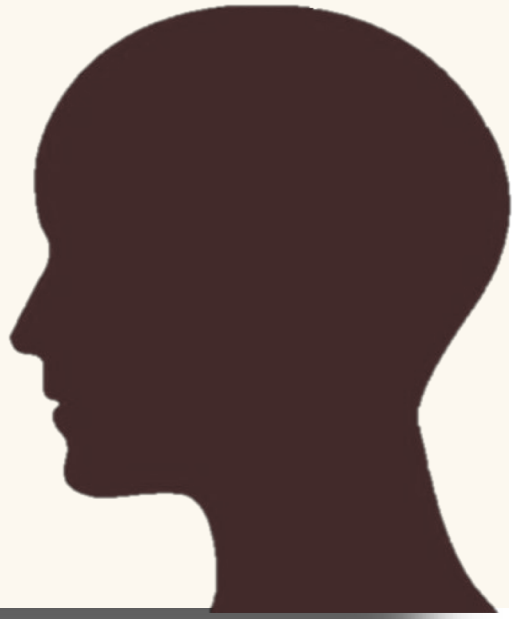
<test audio at 25,000 Epoch>

“????? ???? ??? ??.”

<test audio at 30,000 Epoch>

“????? ???? ??? ??.”

<test audio at 35,000 Epoch>



5. Conclusion

Train set에서 만족할만한 결과를 얻음

Test set에서 첫/끝음을 잡는데에 성공

→ 하지만 아직 일반화가 불가능, 학습데이터에서도 만족스럽지 않음

어마어마한 시간이 소요됨 : 파라미터 튜닝의 결과를 확인 할 수 없었음

실제 음성데이터 처리에서 가장 중요한 부분인 잡음 제거, trimming에 대한 학습을 진행하지 못함

→ 신경망 모형을 다루기 위해 가장 중요한 것은 이해도가 아닌 컴퓨터,,,,

5. Conclusion

이번 프로젝트를 통해 얻은 것!

신경망 모형의 구조와 간단한 신경망의 구현

다양한 라이브러리의 사용법

내장 함수의 중요성 인지

줄어들은 배터리 수명, 컴퓨터에 남아있는 잔열, 4GB에 달하는 모델 체크 포인트 등„


Do you
have any
question?
Thank you
for your attention.