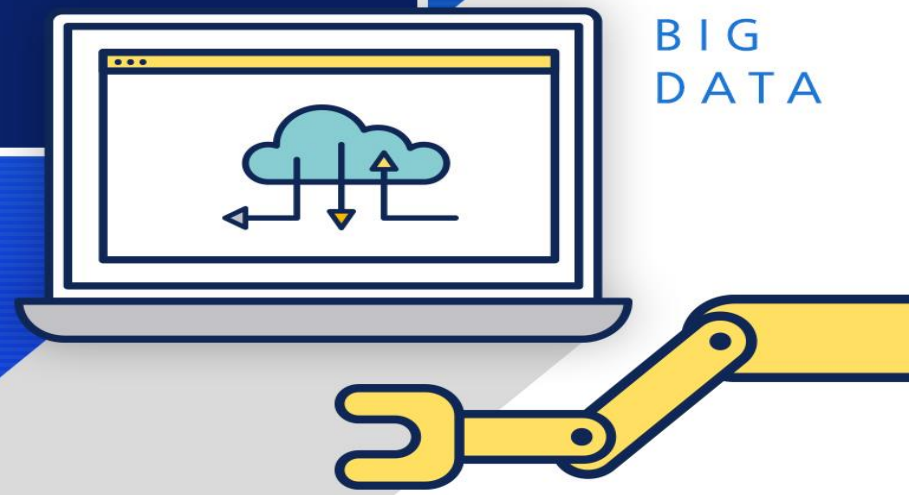
multicampus

혁신성장 청년인재 집중양성 2기

MOF

전효진 김현아 박솔 서상원 이혜민 장동현



Speech To Speech

MOF

Make Or Fake

Members

Hyojin Jeon
Hyeona Kim
Donghyun Jang
Sangwon Seo
Hyemin Lee
Sol Park

Modelling

Hyojin Jeon
Hyeona Kim
Donghyun Jang

UI

Sangwon Seo
Donghyun Jang

Preprocessing

Hyojin Jeon
Hyeona Kim
Donghyun Jang
Sangwon Seo
Hyemin Lee
Sol Park

Text
To
Speech

Speech
To
Speech

Speech
To
Text

기획 배경

의사 / 힘 (X) ▶ **엘사 / 힘 (O)**

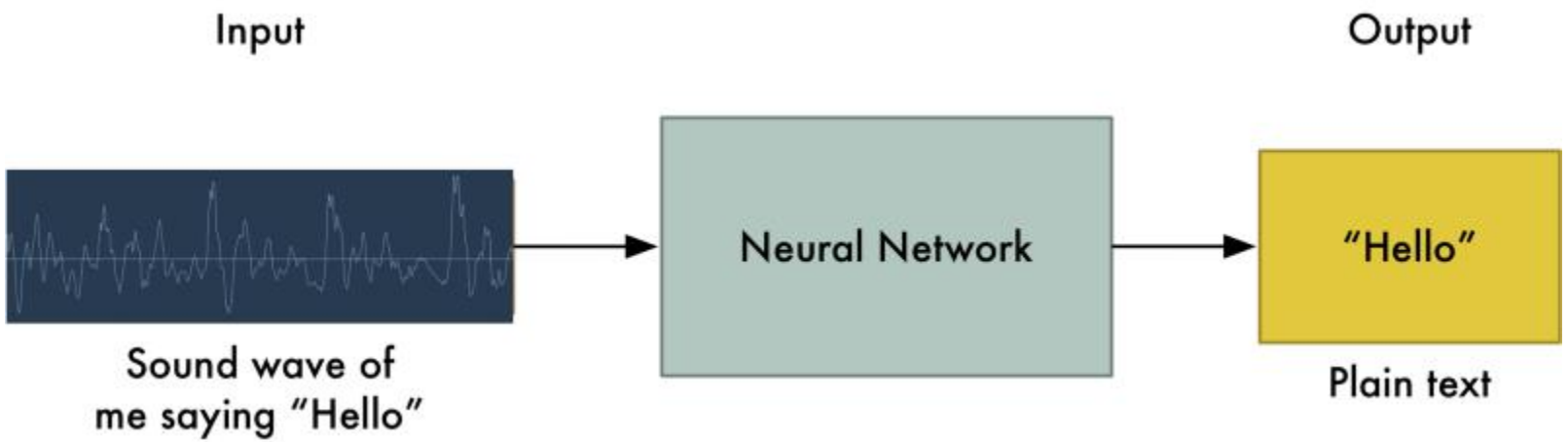
보다 더 **정확한** 자막을 구현할 수는 없을까?

개발환경

	STT	TTS
CPU	Intel Xeon CPU E5-2686 v4 @ 2.30GHz	
RAM	64GB	
GPU	Tesla K80	
OS	Ubuntu 16.04.6 LTS	
Python	3.6	
Cuda	10.0	8.0
Tensorflow	1.15.0rc3	1.3

Speech To Text

STT



"Hello"

"Heeeellooooooo"

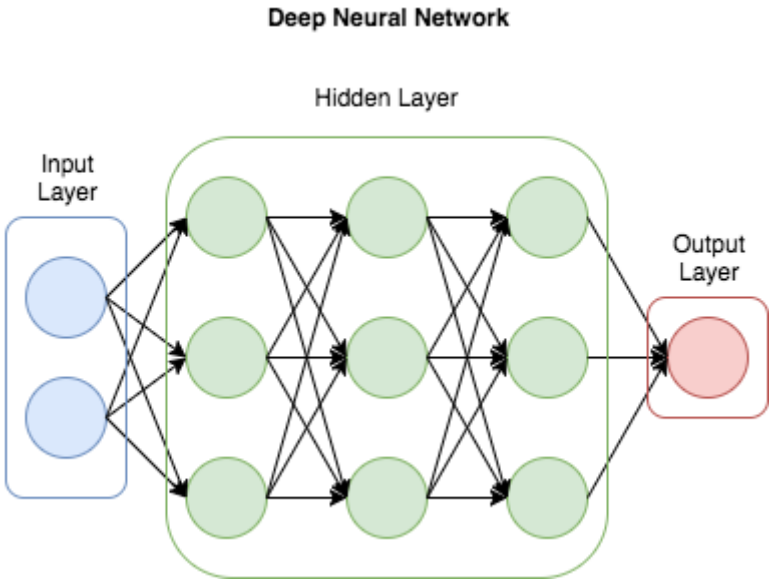
"Hello"

동일한

텍스트로 인식

STT

너무 어려워요



KALDI

Korean ASR Model

Kaldi-based Acoustic Model Design

- TDNN (with Factorization) / TDNN + LSTM / TDNN + OPGRU
- Chain model
- Data augmentation of reverberant speech

Data-driven Language Model Design

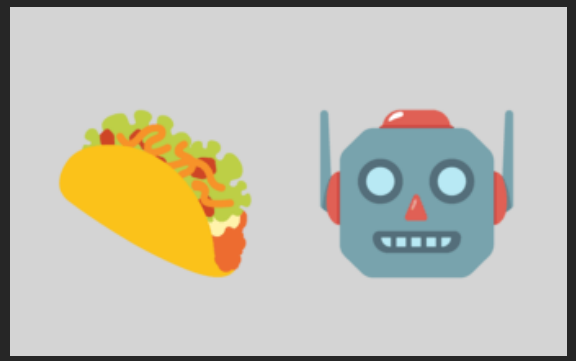
- Text normalization
- Data-driven morpheme analysis
- Automatic Lexicon building
- N-gram Language model

Validations & pre-processings

Audio/Text Crowdsourcing through **MoreCoin** App.

Text To Speech

TTS



Tacotron 을 사용하여 개발

▼

아마존 AWS 서버에서

비 정제 데이터 총 460,000번 + 정제 데이터 기 학습모델에 전이학습으로 2일 학습

Preprocessing

Google Cloud
Speech to Text

연속된 음성 파일을 한 문장열의 형태로 음성간 묵음 기준으로 **분할**

분할된 음성파일을 구글 STT API를 통해 **텍스트 추출**

전처리 과정을 통해 학습에 알맞은 내용만 **선별**

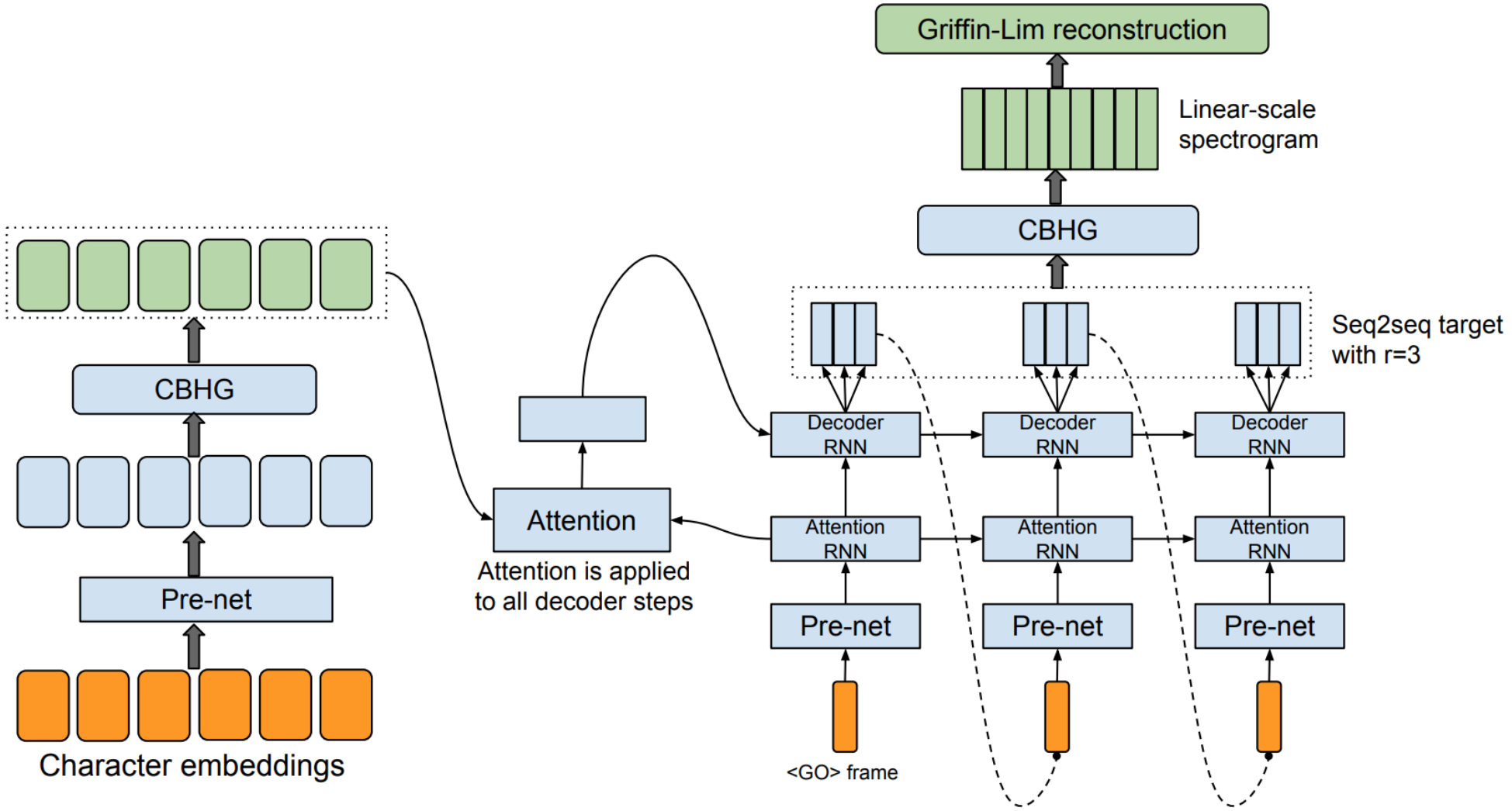
Model - Tacotron

Encoder

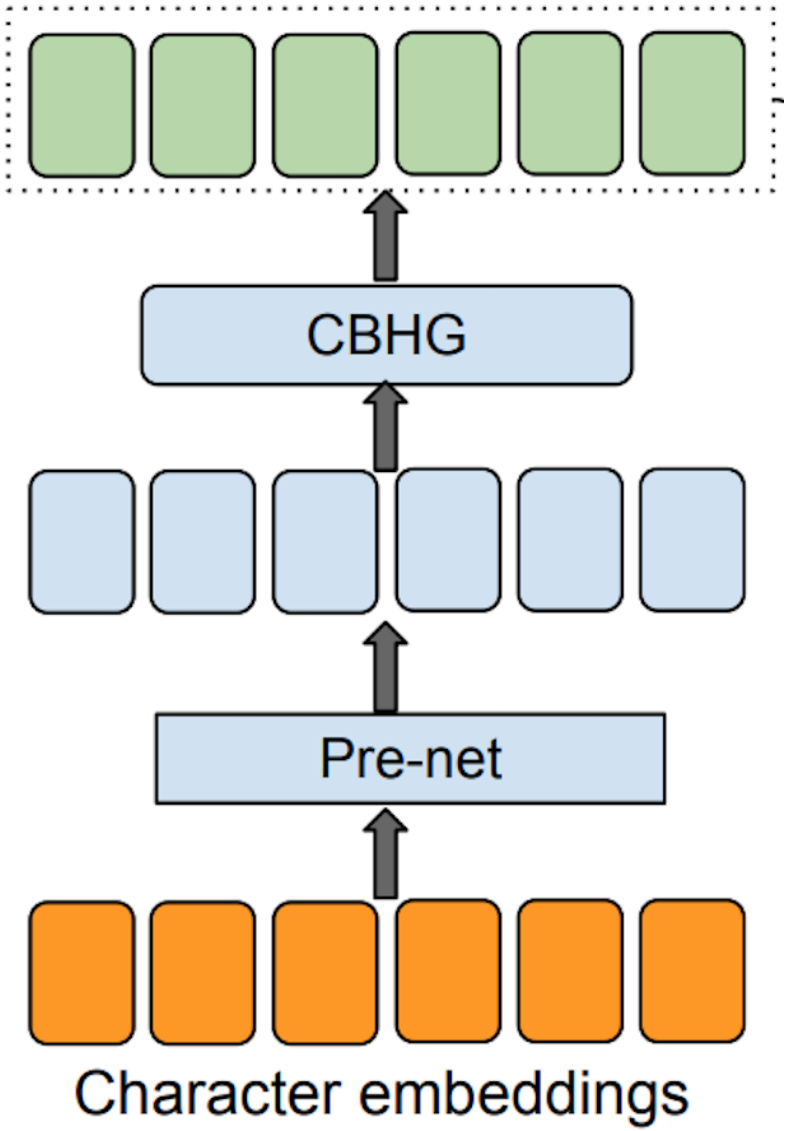
Decoder

Vocoder

Attention



Tacotron - Encoder



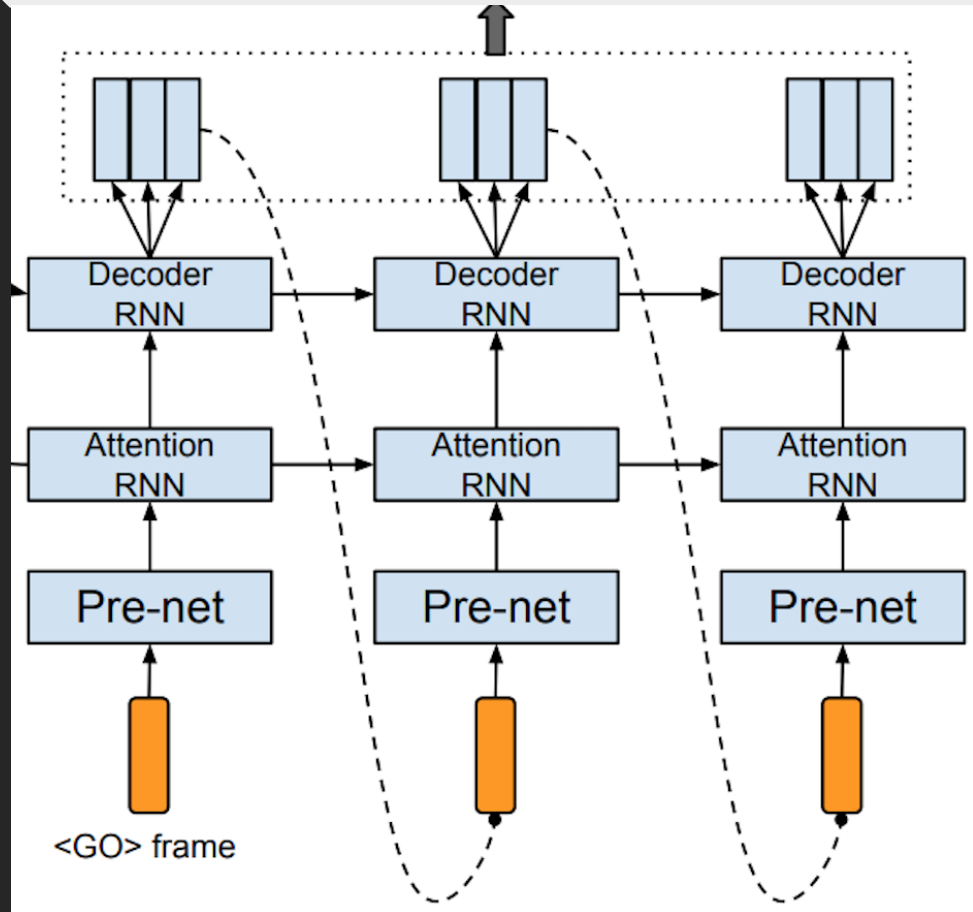
Encoder 의 역할:

텍스트 임베딩 (텍스트를 잘 나타내는 숫자)

딥러닝 모델이 글자를 계산 할 수 없으므로 **숫자**로 바꿔준다

띵작을 말하고 싶을 때, 학습 데이터에 **띵작**이라는 단어가 없어도
명**작**, **띵**동 이라는 단어를 캐릭터 임베딩으로 잘 배웠다면,
명작을 잘 표현하는 캐릭터 임베딩을 만들 수 있다

Tacotron - Decoder

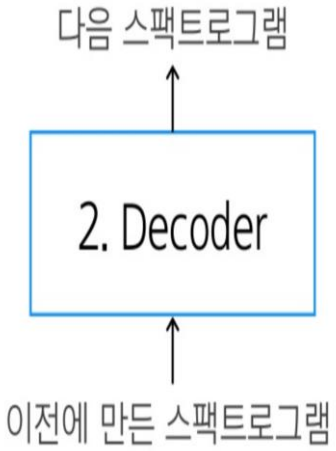


처음에 입력으로 〈GO〉 frame을 받는데.
이것은 학습되는 어떤 숫자들이다

Decoder 의 역할:

n개의 스팩트로그램의 리스트를 얻게 한다

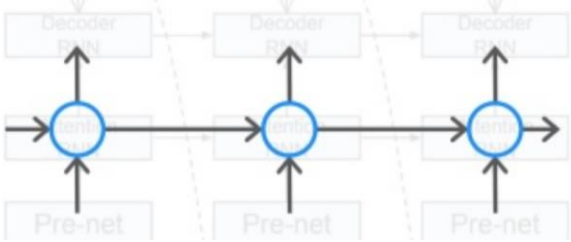
: 음성이 되기 직전의 숫자들



이전에 만든 스팩트로그램을 입력으로 받아

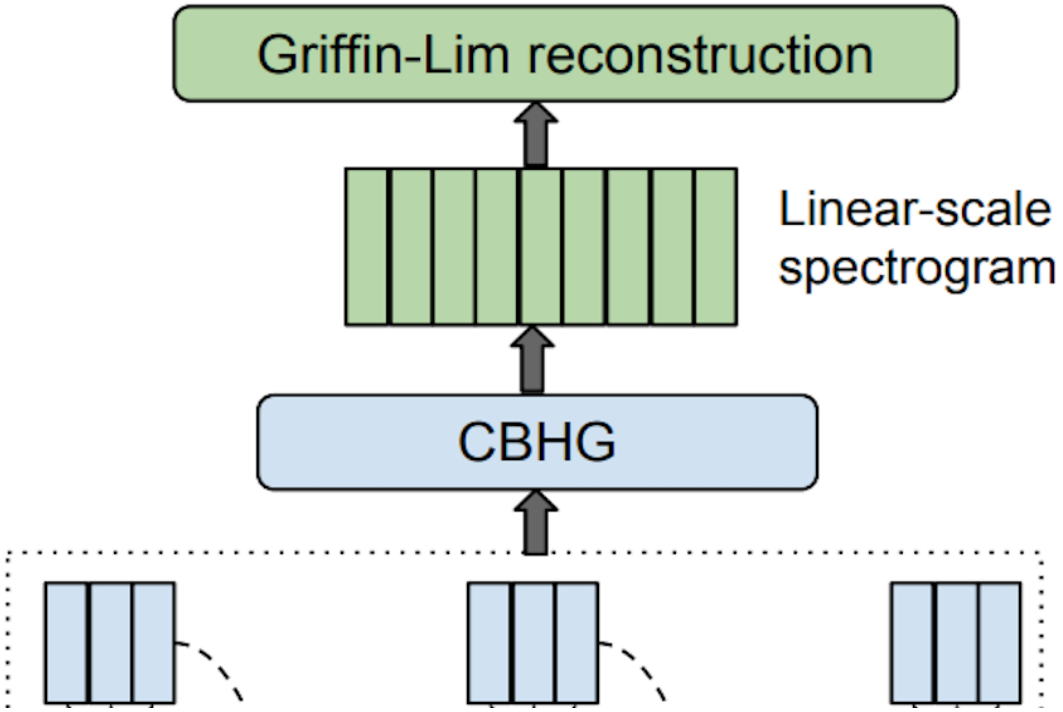
Decoder로 지나고

다음 스팩트로그램을 생성한다



크게 보면 RNN 구조이다

Tacotron – Vocoder / Attention

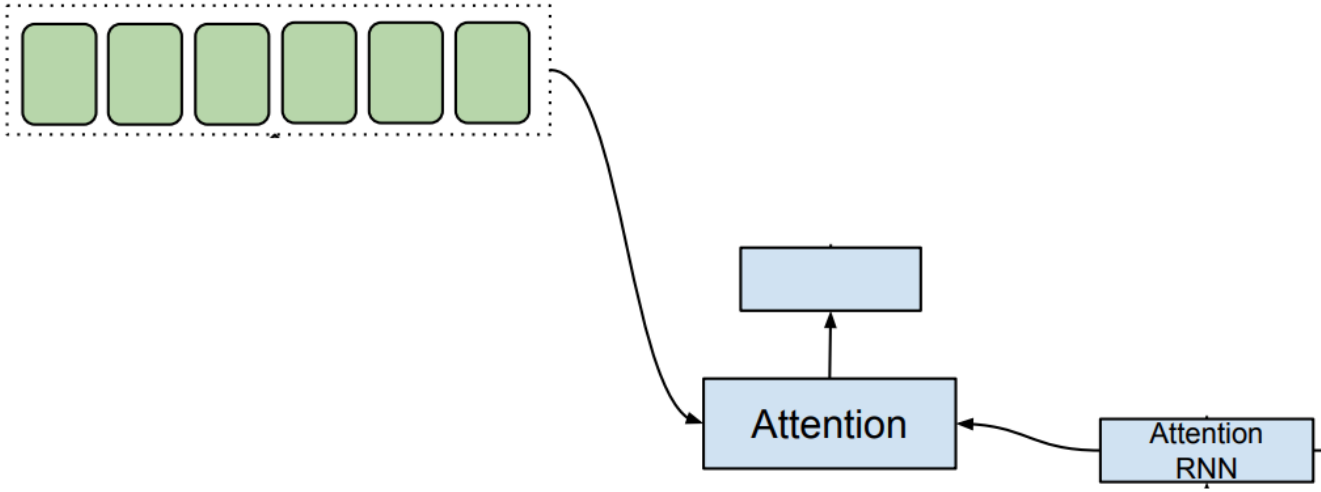


Vocoder 의 역할:

스팩트로그램을 입력으로 받아서

그리핀-림(Griffin-Lim) 알고리즘을

이용하여 음성을 생성한다



Attention:

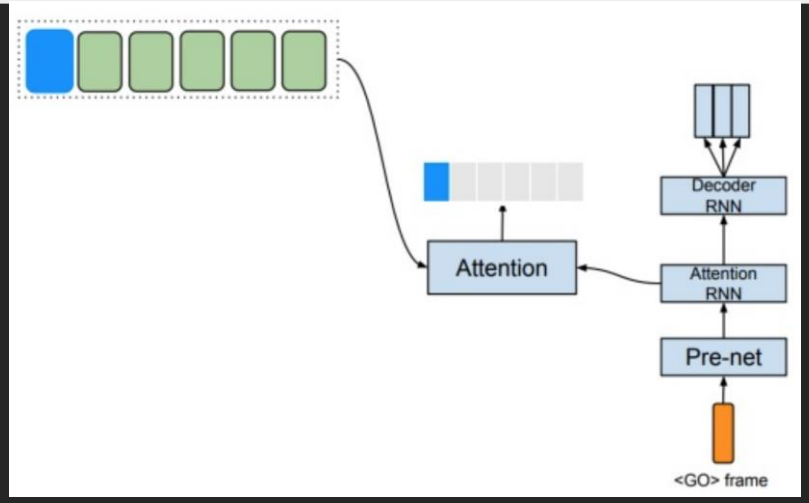
어디에 집중할 것인가?

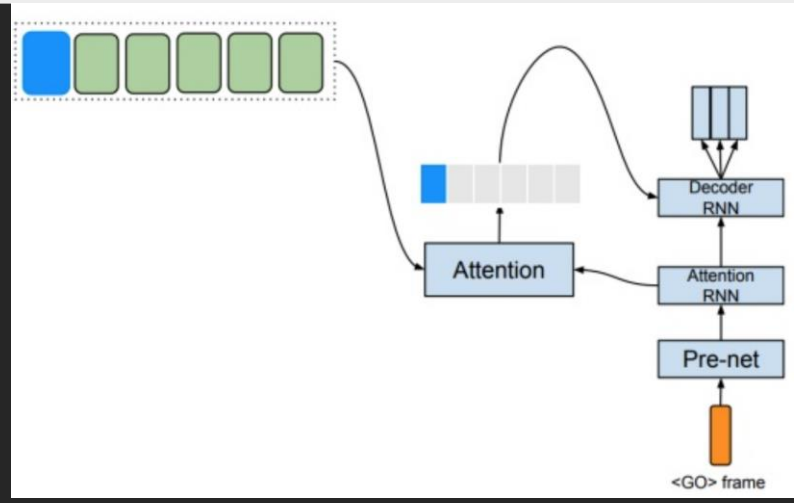
텍스트 임베딩과 Decoder를 잘 합쳐주는 모델

왼쪽 초록색 배열은 Encoder의 출력

= 텍스트 임베딩 한 결과물

Tacotron - Attention





첫번째 음성은
텍스트의 첫번째 단어에 집중

어디에 집중할 것인지
학습을 통해 알아서 계산

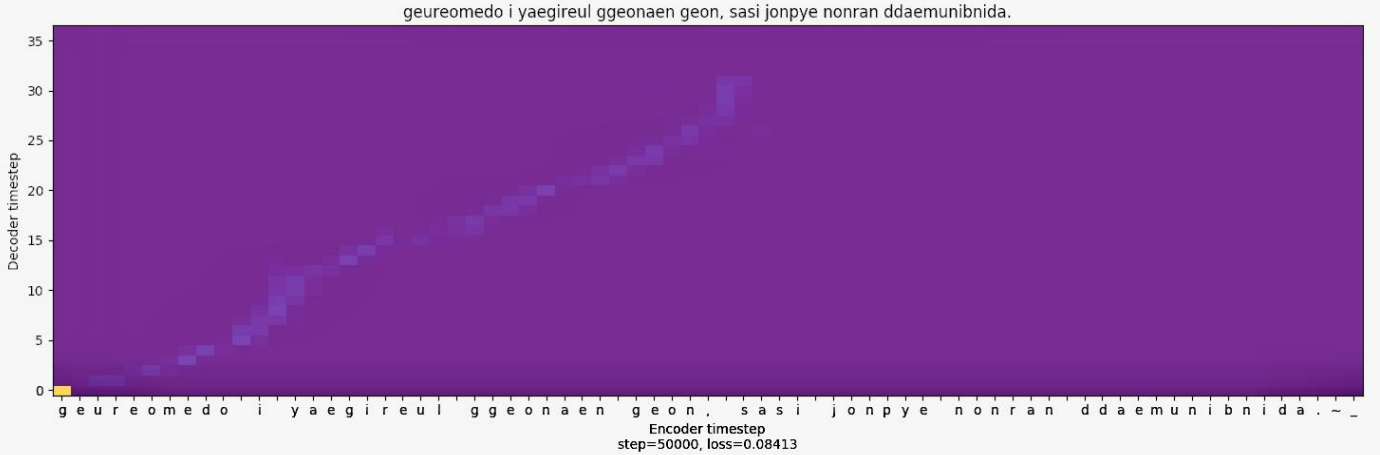
스팩트로그램을 만드는
RNN에 Attention을 전달

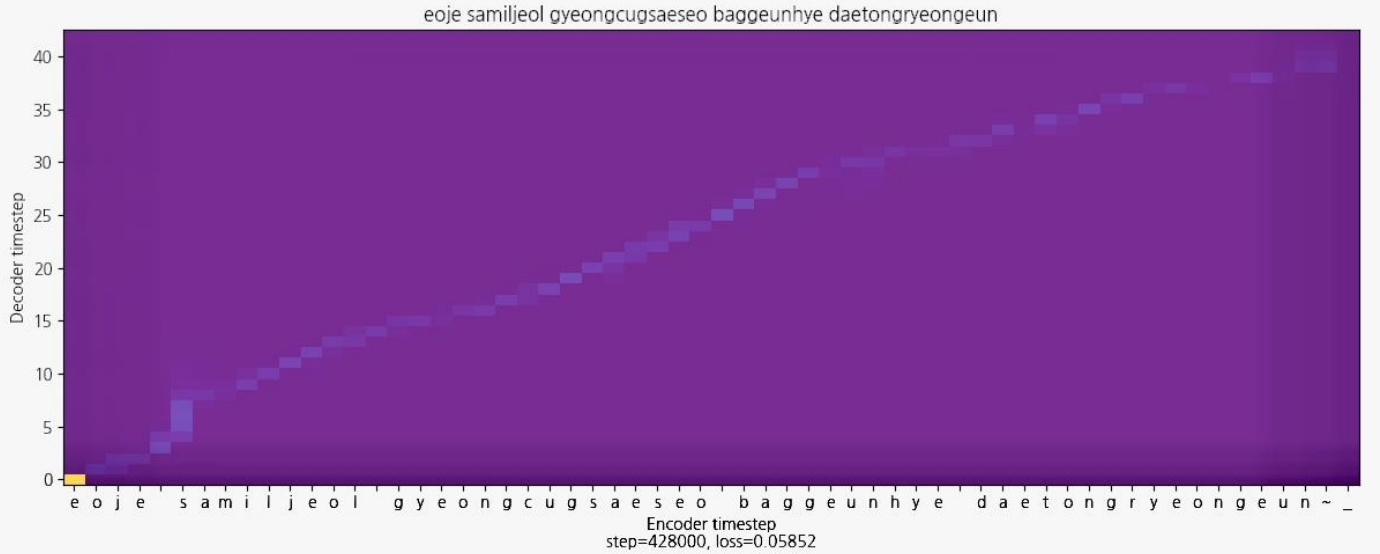
Attention이 중요한 이유

일반화의 중요성 = 학습하지 않았던 문장도 얼마나 잘 말할 수 있는가

TTS

대표적 parameter 값	
cleaners	korean_cleaners
model_type	single # [single, simple, deepvoice]
attention_size	f(256)
batch_size	32
initial_learning_rate	0.002
reduction_factor	4
min_iters	20
max_iters	200





3

Speech To Speech

Thank you

MOF