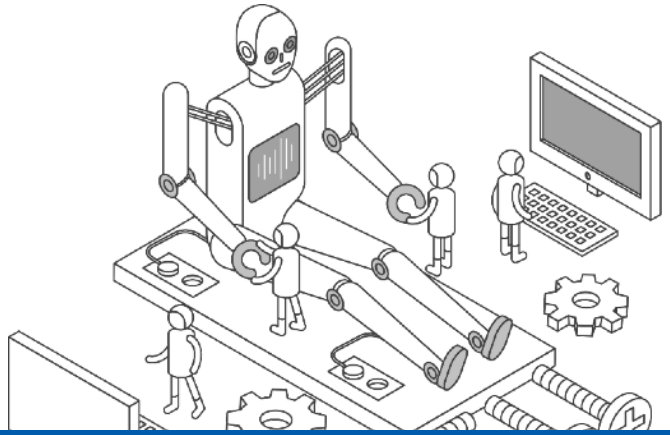
2021 직업계고 AI 전문교육

ARTIFICIAL INTELLIGENCE
BIG DATA
SMART FACTORY

AI·빅데이터 심화과정

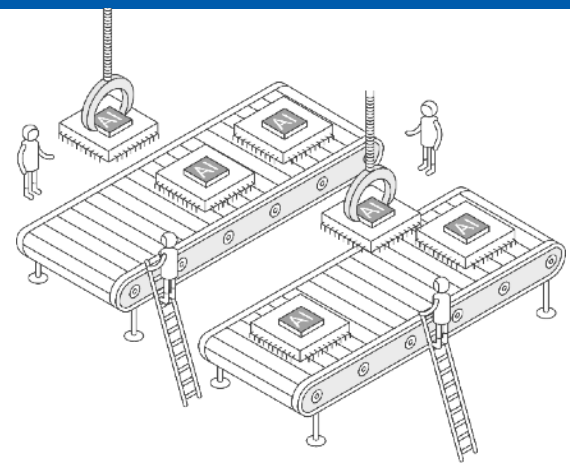
# Tacotron2: 딥러닝으로 만드는 TTS

**박수철**
github.com/scpark20
GaudioLab, 모두의연구소



서울특별시교육청 kocoafab (주) 네 패 스 코 코 아 랩 모두의 연구소

# Tacotron2: 딥러닝으로 만드는 TTS



# 후반기 일정

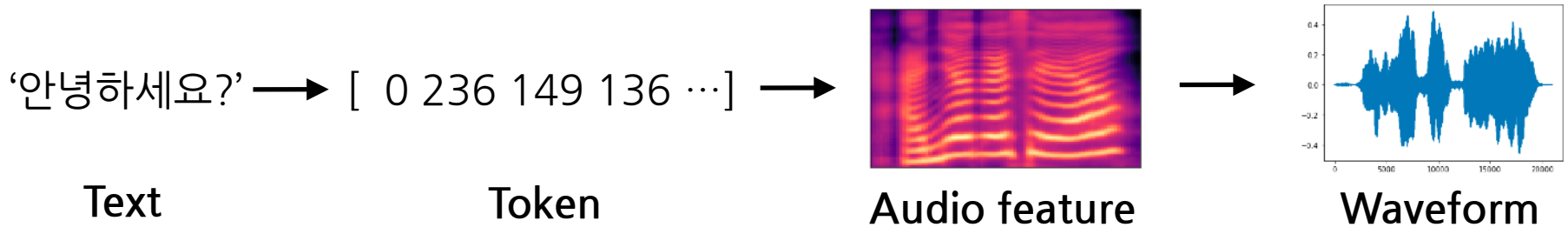
**Seq2Seq**
9/24 - RNN으로 소설쓰기 (Aiffel 외)
9/29, 10/1 - 26. 뉴스 요약봇 만들기
10/6, 10/8, 10/15 - 27. 트랜스포머로 만드는 대화형 챗봇

**CNN/GAN**
10/20, 10/22 - 22. 난 스케치를 할테니 너는 채색을 하거라
10/27, 11/3 - 21. 흐린 사진을 선명하게

**RNN+CNN**
11/5 - RNN으로 음성인식하기 (Aiffel 외)
11.10 - 19. 직접 만들어보는 OCR

**Text-to-Speech**
**11/12, 17 - Tacotron2: 딥러닝으로 만드는 TTS**

- TTS는 Text-to-Speech의 약자로 주어진 문장을 음성 데이터로 변환하는 작업을 말하며, speech synthesis라고도 말합니다.
- 어진 음성 데이터를 문장으로 변환하는 작업인 ASR(Automatic Speech Recognition)와 정확히 반대 개념이라고 생각할 수 있습니다.
- Audio feature는 현재에는 mel-spectrogram이나 waveform을 주로 사용하고 과거에는 MFCC를 사용했습니다.
- 입력이 되는 token들은 다음과 같이 integer들의 sequence로 표기할 수 있습니다.
  $\mathbf{x} = (x_1, x_2, \ldots, x_L)$, where $x_l \in \mathbf{N}$
- 출력이 되는 audio feature들은 다음과 같이 d-dimensional vector들의 sequence로 표기할 수 있습니다.
  $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T)$, where $\mathbf{y}_t \in \mathbf{R}^d$



- Tacotron2는 2018년에 Natural tts synthesis by conditioning wavenet on mel spectrogram predictions 논문에서 제안되었습니다.
- Tacotron2은 크게 text를 인코딩하는 encoder와 mel-spectrogram을 디코딩하는 decoder 모듈로 이루어져 있습니다.
- 이에 더해 attention 모듈이 encoder와 decoder 모듈간의 정보 전달을 위해 사용됩니다.
- 또한 별도의 모델인 WaveNet을 통해 mel-spectrogram을 waveform으로 변환하여 사실적인 음성을 만드는데 기여합니다.

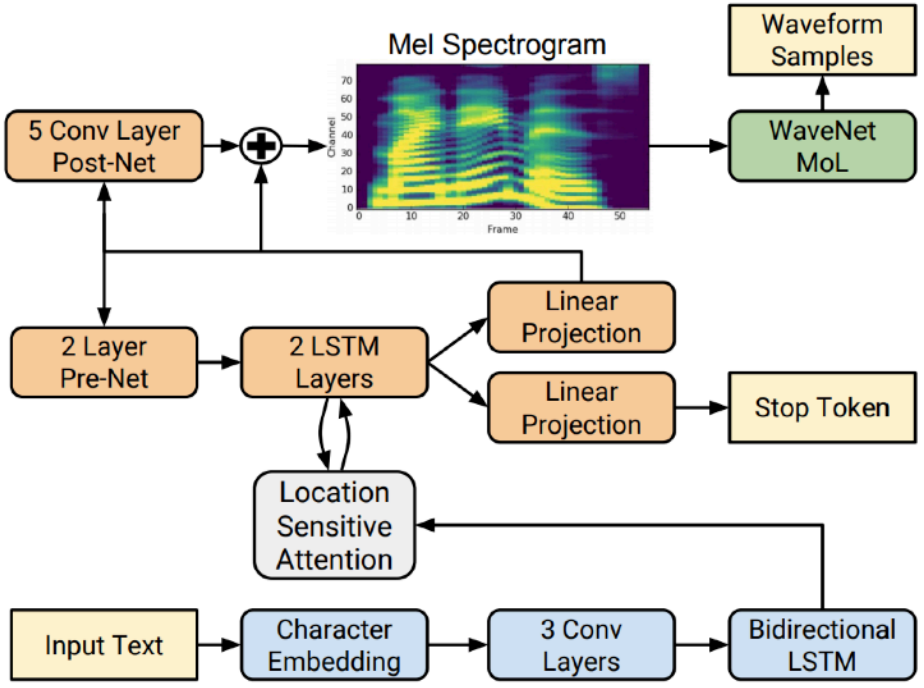



**Fig. 1.** Block diagram of the Tacotron 2 system architecture.

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

- Encoder는 input text를 입력받아 decoder에 넘겨줄 encoding data를 만드는 역할을 합니다.
- Character embedding : input text (tokens)를 입력받아 embedding table을 이용하여 vector sequence를 출력합니다. Embedding table은 trainable한 parameter로 이루어져 있습니다.
- 3 Conv Layers : 가까운 거리에 있는 token간의 context를 파악하는데 사용합니다.
- Bidirectional LSTM : 멀리 떨어져 있는 token간의 context를 파악하는데 사용합니다.

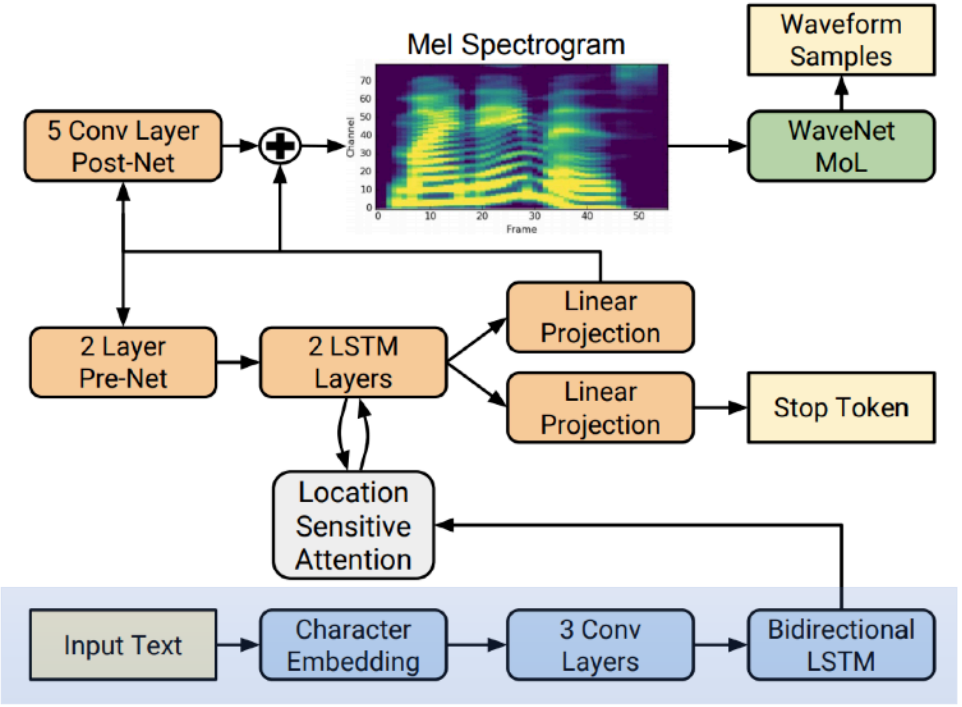



**Fig. 1.** Block diagram of the Tacotron 2 system architecture.

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

- Decoder는 auto-regressive하게 작동합니다. 즉, 이전의 mel frame들을 받아 현재 시점의 mel frame을 생성합니다. 이 때, attention 모듈에 의해 encoding data를 참조합니다.

- 2 Layer Pre-Net : 이전 step의 mel frame을 받아 두 개의 linear layer를 적용한 결과를 출력합니다.

- 2 LSTM Layers : 이전의 mel frame들의 context를 분석하는데 사용합니다. 두 LSTM 사이에 attention 모듈에 의해 가져온 context를 concat으로 더합니다.

- Linear Projection : mel frame과 stop token을 얻는데 사용합니다.

- 5 Conv Layer Post-Net : Auto-regressive하게 한 frame씩 생성한 mel-spectrogram의 품질을 향상시키기 위해 사용합니다.

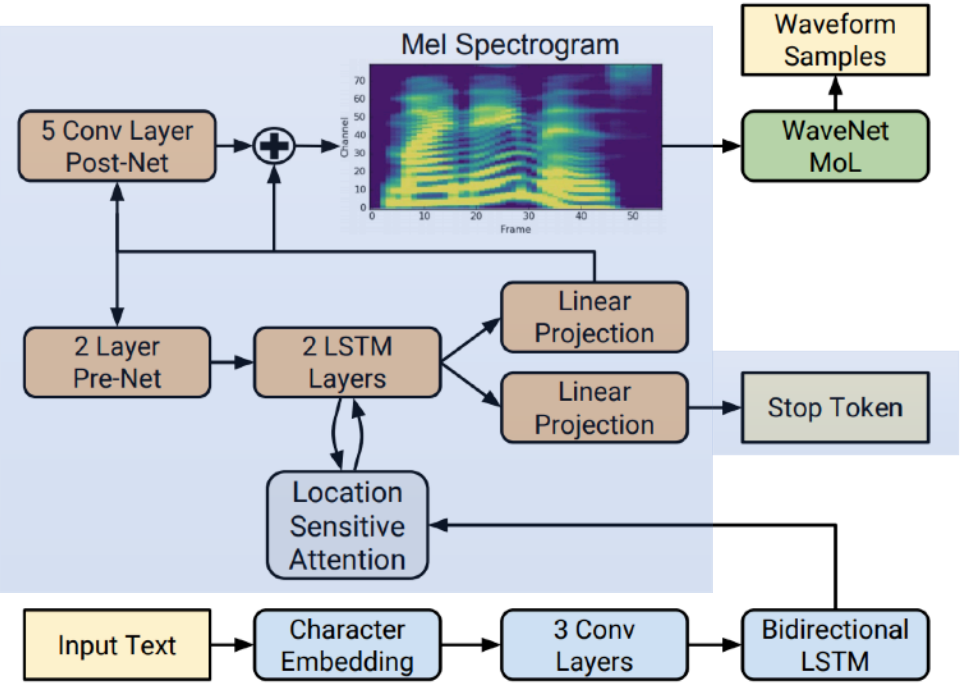



**Fig. 1.** Block diagram of the Tacotron 2 system architecture.

Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.


똑똑
파이팅!!