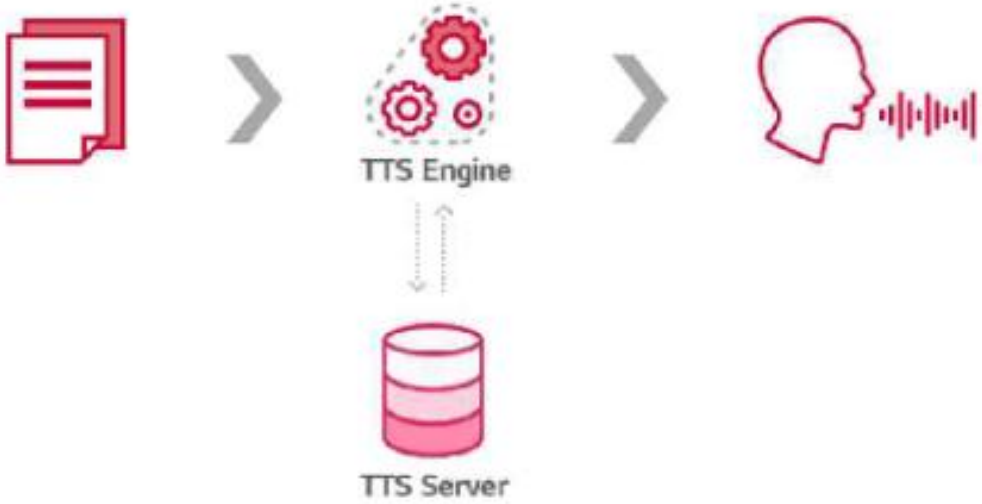
Part 1

TTS 시스템



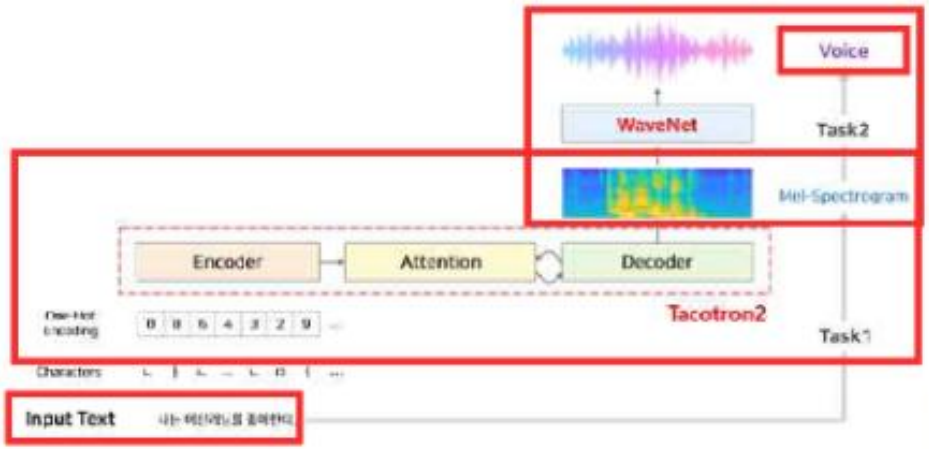
TTS - Text To Speech

- 텍스트를 입력하면 목소리로 변환해주는 기술
- 기계와 인간이 대화할 수 있도록 하기 위한 기초기술
- 영어, 숫자, 한글 등의 문자를 입력하면 자연스러운 인간의 음성으로 출력

TTS 는 Text To Speech 의 약자로 텍스트를 입력하면 목소리로 변환해주는 기술이며 음성합성 기술에 속합니다.

기계와 인간이 서로 대화할 수 있도록 하기 위한 기초기술로 영어, 숫자, 한글 등의 문자를 입력하면 자연스러운 인간의 음성으로 출력해줍니다.

Part 1 TTS 시스템



- 모델은 텍스트를 받아 음성을 합성
- Input: 텍스트(text)
  Output: 음성 (voice)

- 텍스트로부터 Mel - spectrogram을 생성하는 단계

- Mel - spectrogram으로부터 음성을 합성하는 단계

이 그림은 `TTS 시스템` 의 전체 구조입니다.

모델은 텍스트를 받아 음성을 합성합니다. 따라서 최종 Input은 텍스트, Output은 음성입니다.

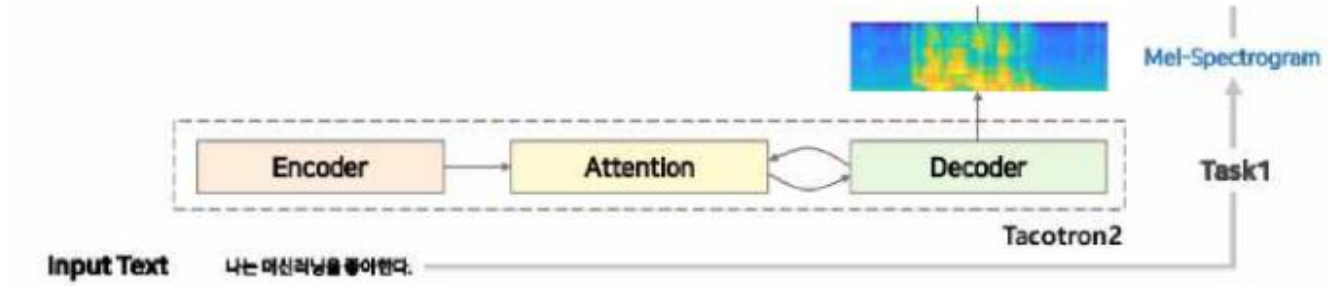
하지만 텍스트로부터 바로 음성을 생성하는 것은 어려우므로 `TTS` 를 두단계로 나누어 처리합니다.

첫 번째 단계는 텍스트로부터 . `Mel-spectrogram` 을 생성하는 단계이며, 두 번째 단계는 `Mel-spectrogram` 으로부터 음성을 합성하는 단계입니다.

여기서 `Mel-spcetrogram` 이란 현재 가지고 있는 주파수를 좀 더 인간의 귀의 특성에 알맞게 주파수를 쪼개서 분석하는 방법 입니다.

밑에 보이는 `encoder` 와 `attention` 과 `decoder` 로 구성된 딥러닝 구조의 타코트론 모델이 첫 번째 단계를 담당합니다.

Part2 Tacotron2 - 개념 및 구조



Tacotron 2

: 음성 합성에 대표적인 모델로 고품질의 음성을 생성할 수 있는 딥러닝 기반 TTS 모델
: input - character / output - Mel-Spectrogram

- Encoder: character로부터 음성에 특징을 추출하는 것
- Attention: 매 시점 Decoder에서 사용할 정보를 추출하고 할당
- Decoder: Attention에서 얻은 정보와 이전 시점에서 생성된 mel-spectrogram을 이용해
  (1) 현재시점의 Mel-spectrogram 을 생성
  (2) 현재시점의 종료확률을 계산
  (3) Mel-spectrogram의 품질을 향상

`타코트론2` 모델이란 고품질의 음성을 생성할 수 있는 딥러닝 기반의 모델입니다.

입력은 텍스트이고 출력은 `Mel-spectrogram` 이며 인코더, 어텐션, 디코더를 통해 입력이 출력으로 변경되는 구조입니다.

첫 번째 단계인 인코더의 역할은 캐릭터로부터 음성에 특징을 추출하는 것입니다.

두 번째 어텐션의 역할은 매 시점 디코더에서 사용할 정보를 인코더에서 추출하고 할당합니다.

마지막으로 디코더의 역할은 어텐션에서 얻은 정보와 이전 시점에서 생성된 `Mel-spectrogram` 을 이용하여 현재 시점의 `Mel-spectrogram` 을 생성하고 종료확률을 계산하여 `Mel-spectrogram` 의 품질을 향상합니다.

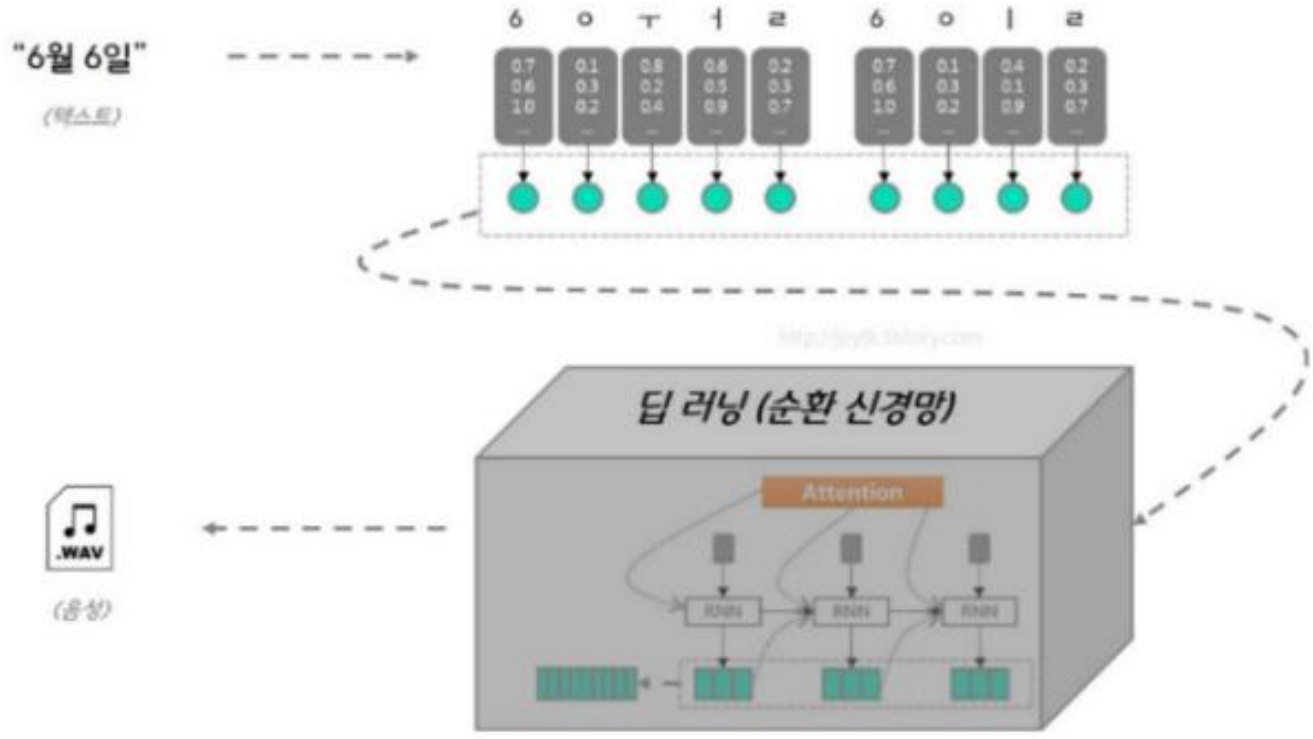
타코트론 학습 방법을 간단하게 설명하자면 다음과 같습니다.

우선 텍스트에서 feature를 추출합니다.

위와 같이 한국어 기준으로 텍스트를 자음 모음 단위로 분리합니다.

그다음 컴퓨터가 쉽게 알아들을 수 있도록 자모들을 숫자로 바꿔줍니다.

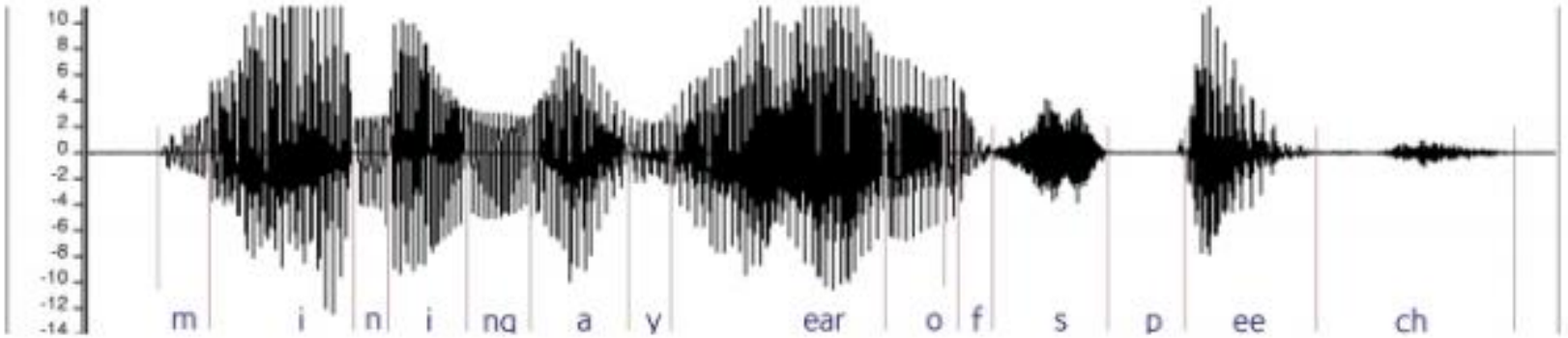
마지막으로 `RNN` 기반의 신경망에서 학습을 시켜 `타코트론` 의 학습을 완료할 수 있습니다.



여기서 `순환 신경망 RNN` 이란 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 인공신경망입니다.

순환 신경망의 대표적인 예로는 글자나 문장을 완성해주는 알고리즘 등이 있습니다.

타코트론2를 시스템에 TTS 시스템에 적용하는 방법



공개된 학습용 TTS 데이터는 KSS . 데이터셋을 활용하려고 합니다.

KSS 데이터셋은 전문 여성 성우 한 분이 한글과 한영사전 4권의 예문을 읽은 약 12시간 분량의 데이터셋입니다.

이 데이터셋을 사전학습한 후 개발자의 녹음된 음성을 이용해 이전에 학습된 모델 가중치로부터 학습을 업데이트할 수 있습니다.