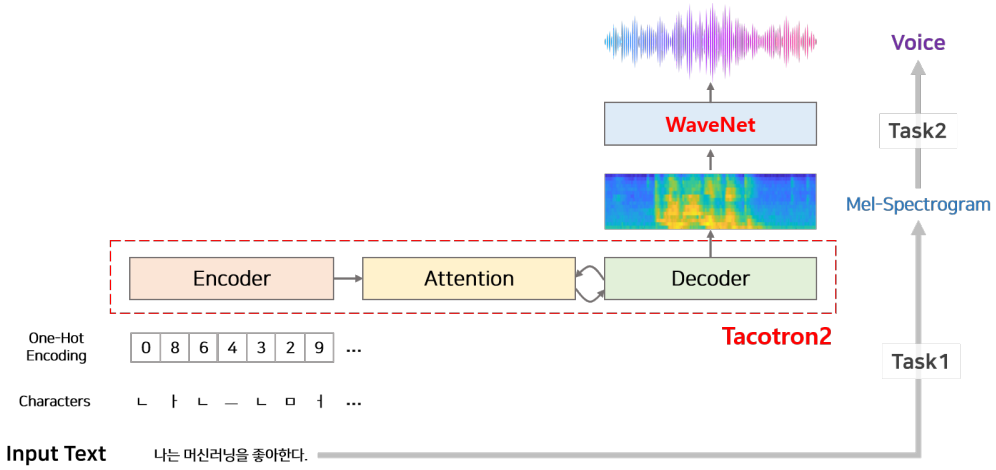
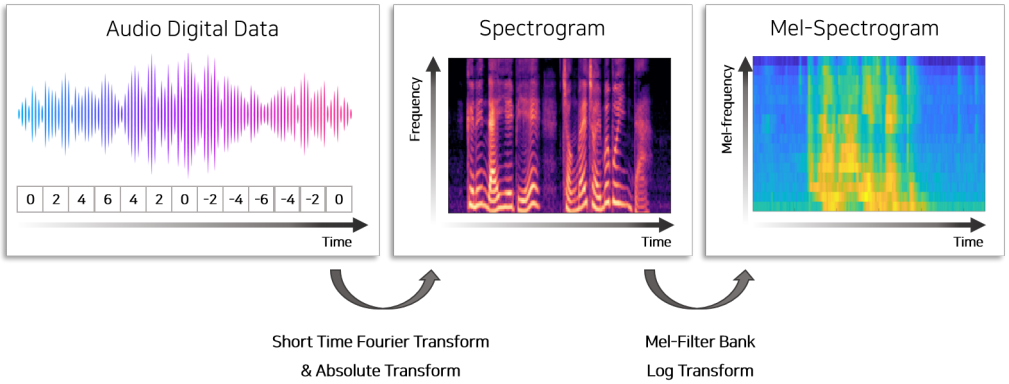
엔비디아 타코트론2



TTS 시스템의 전체 구조입니다.

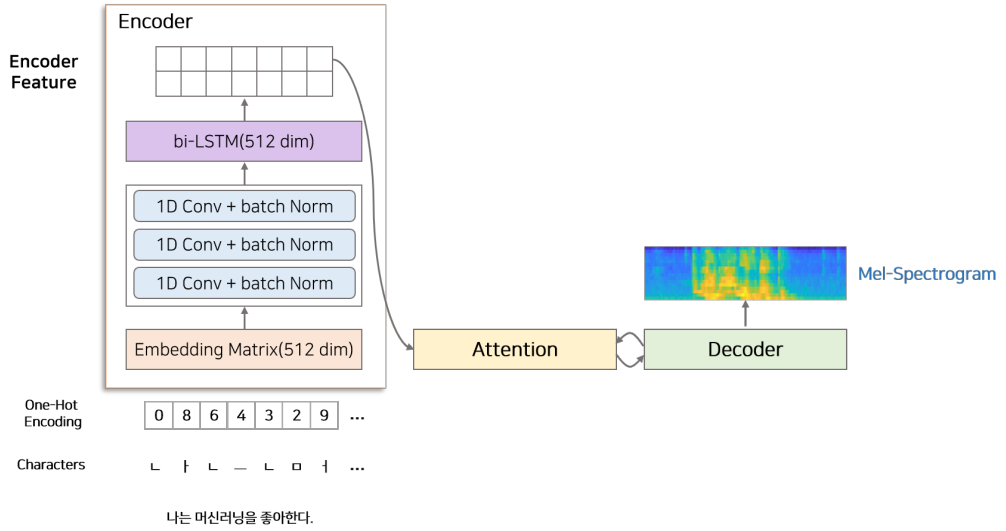
모델은 특정 텍스트를 받아 음성을 합성하며 따라서 최종 입력과 출력은 텍스트와 음성입니다. 하지만 텍스트에서 음성을 바로 생성하는 것은 어려운 과정이므로 TTS를 두 단계로 나누어 처리할 수 있습니다. 첫 번째 단계는 텍스트로부터 Mel-spectrogram을 생성하는 단계이며 두 번째 단계는 Mel-spectrogram으로부터 음성을 합성하는 단계로 나누어집니다. 여기서 인코더와 디코더로 구성되어있는 딥러닝 구조의 타코트론2 모델이 첫 번째 단계를 담당합니다.

타코트론2 모델의 input은 텍스트를 자음과 모음 단위로 쪼갠 character이고 output은 mel-Spectrogram입니다. 모델은 크게 Encoder, Decoder, Attention 모듈로 구성되어 있습니다. Encoder는 character를 일련 길이의 hidden 벡터(feature)로 변환하는 작업을 담당합니다. Attention은 Encoder에서 생성된 일정길이의 hidden 벡터로부터 시간순서에 맞게 정보를 추출하여 Decoder에 전달하는 역할을 합니다. Decoder는 Attention에서 얻은 정보를 이용하여 mel-spectrogram을 생성하는 역할을 담당합니다.

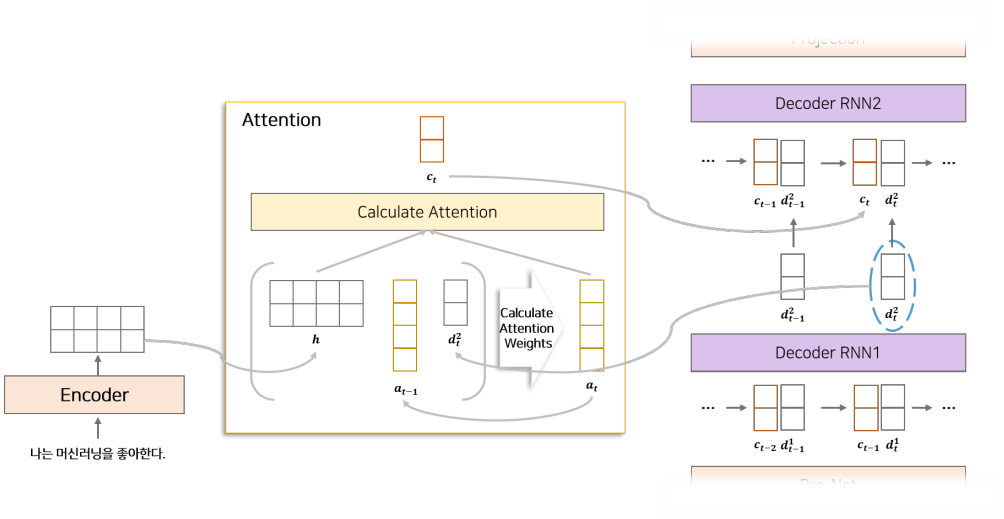


이 모델을 학습하기 위해서 입력과 출력이 한쌍으로 묶인 데이터가 필요합니다. 텍스트와 음성이 쌍으로 묶인 데이터가 있다면 이를 모델의 input과 output(label)의 형태로 가공하여야 합니다. 즉 텍스트는 character로 만들어야 하고 음성은 Mel-spectrogram으로 변형해야 합니다. 예를 들어 텍스트가 “나는 머신러닝을 좋아한다.” 라면 ‘ㄴ’, ‘ㅏ’의 character 형식으로 변경됩니다. 이후 인코딩 과정을 통해 정수열로 변경한 뒤 모델의 입력으로 활용합니다.

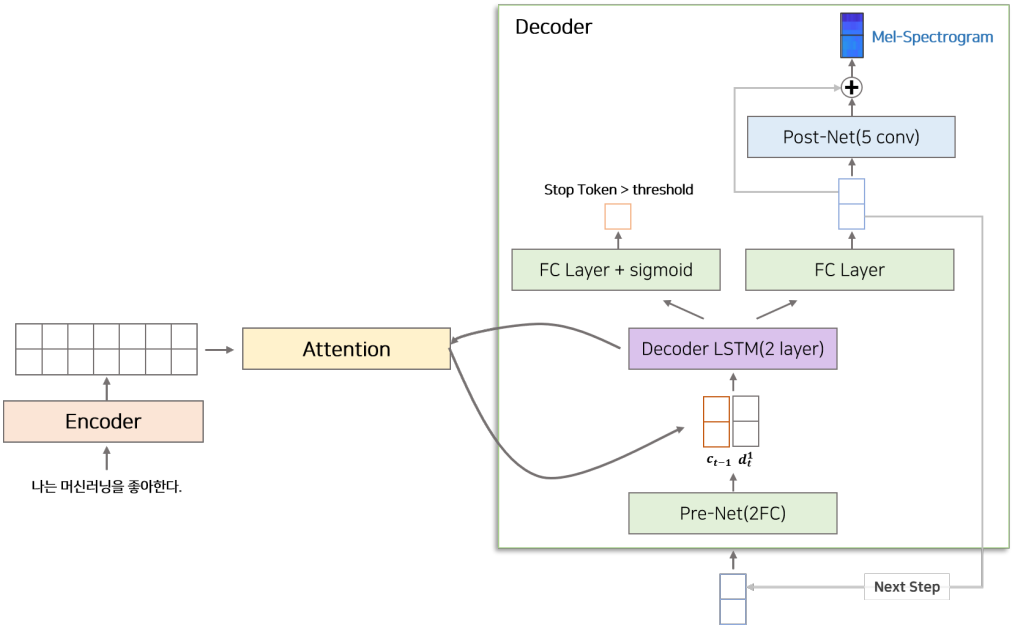
음성데이터로부터 mel-spectrogram을 추출하여 출력을 활용하기 위해 전처리 작업이 필요합니다. 첫 번째는 여러 개의 오디오, 즉 주파수가 섞여 있는 오디오데이터의 오디오를 분리해 표시하기 위해 푸리에 변환을 활용합니다. 두 번째로 저주파 영역을 확대하는 작업을 수행합니다. 사람의 귀는 고주파보다 저주파에 더 민감하므로 저주파 영역을 확대하고 고주파의 영역을 축소하여 사용합니다.



Encoder 과정은 character 단위의 one-hot vector를 encoded feature로 변환하는 역할을 합니다. Character Embedding, 3 Convolution Layer, Bidirectional LSTM으로 구성되어 있습니다.



Attention은 매 시점마다 Decoder에서 사용할 정보를 Encoder에서 추출하여 가져오는 역할을 합니다. 즉 Attention 메카니즘은 인코더의 LSTM에서 생성된 feature와 Decorder의 LSTM에서 전 시점에 생성된 feature를 이용하여 인코더로부터 어떤 정보를 가져올지 정렬하는 과정을 의미합니다.



Decorder는 Attention 단계를 통해 얻은 정렬 feature와 이전 시점에서 생성된 mel-spectrogram 정보를 이용하여 다음 시점의 mel-spectrogram을 생성하는 역할을 합니다. Decorder는 Pre-net, Decorder LSTM, Projection Layer, Post-Net으로 구성됩니다.

Pre-Net은