End-to-End ASR
: Deep Speech 2

MILab Undergraduate student, Kim Taehyeon

2023. 04. 21

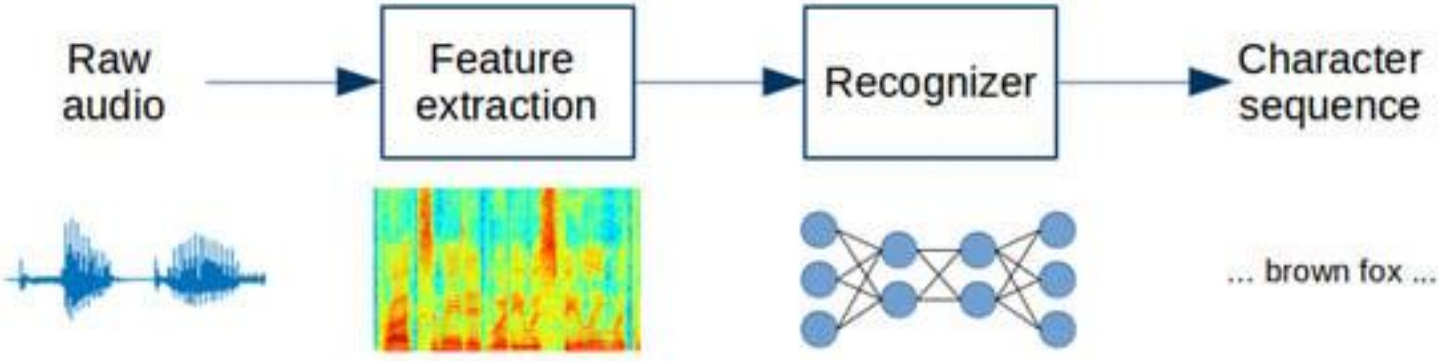
목차

1. 음성인식 학습 모델 종류

1. 음성인식 학습 모델 종류

1. End-to-End ASR(Automatic Speech Recognition)

End-to-End ASR 은 입력되는 음성을 받아서 처리하고 최종적으로 이를 문자로 표현하는 것이다.



1. 음성인식 학습 모델 종류

1. 음성인식 학습 모델 종류

2. Keyword Spotting

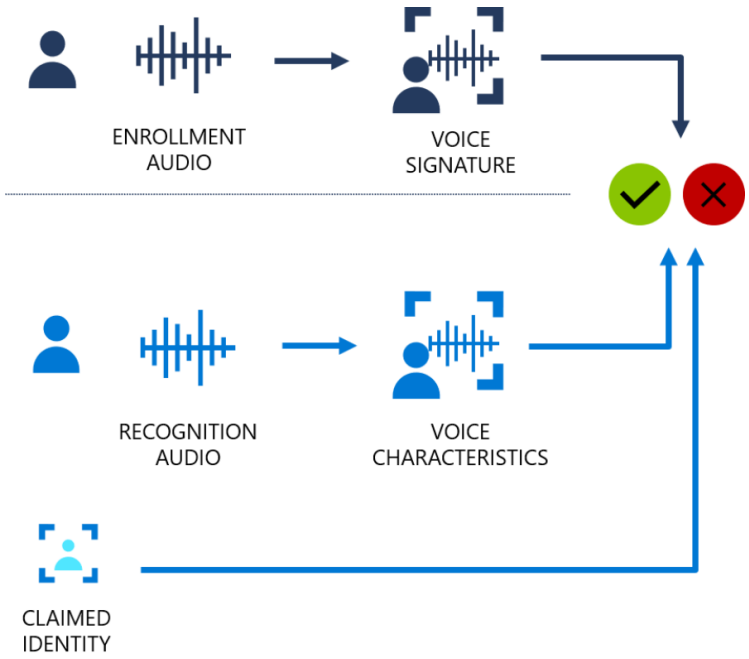
Keyword Spotting 은 대표적인 AI 비서인 KT Genie 의 "지니야", Apple Siri 의 "Hey, Siri" 와 같은 호출어를 인식하도록 하는 것이다.

1. 음성인식 학습 모델 종류

1. 음성인식 학습 모델 종류

3. Speaker Recognition

사람마다 고유한 음색이 존재한다는 사실을 바탕으로 음성 데이터를 활용하여 사용자 판별 및 사용자 별 맞춤 서비스를 제공할 수 있다.

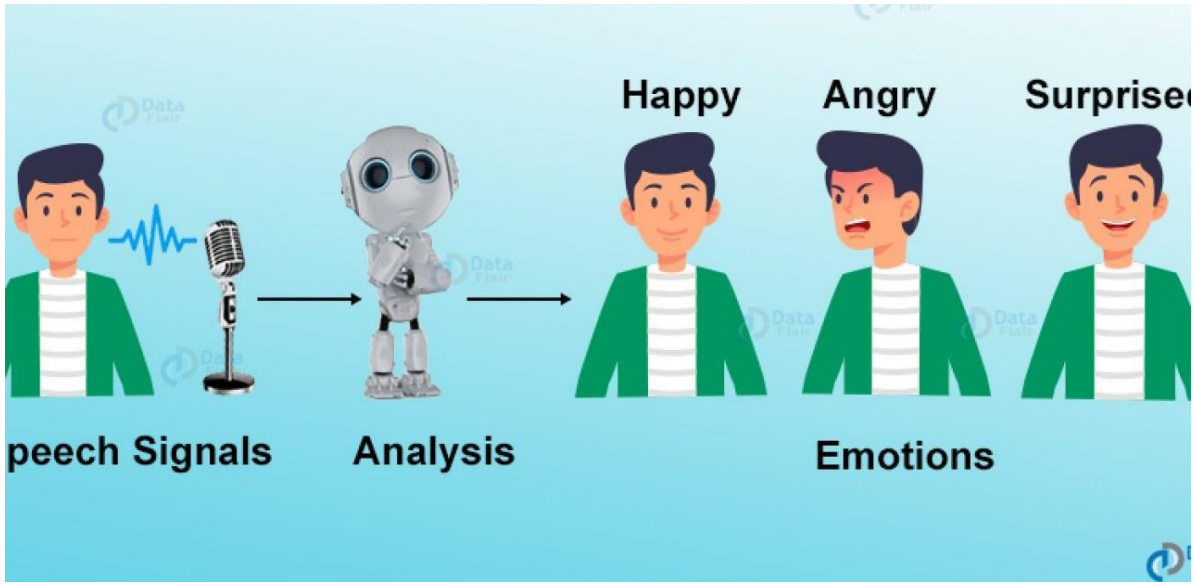


1. 음성인식 학습 모델 종류

1. 음성인식 학습 모델 종류

4. Emotion Recognition

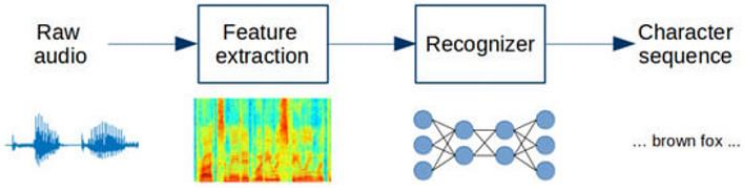
목소리의 변화를 감지하여 사용자의 감정을 파악할 수 있는 방식이다.



2. End-to-End ASR 딥러닝 모델 원리

1. End-to-End ASR(Automatic Speech Recognition)

End-to-End ASR 은 입력되는 음성을 받아서 처리하고 최종적으로 이를 문자로 표현하는 것이다.

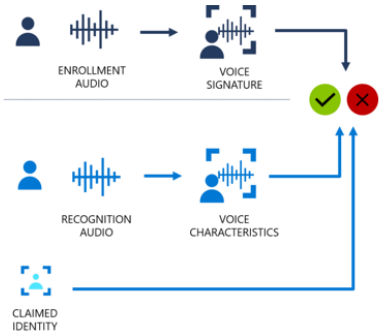


2. Keyword Spotting

Keyword Spotting 은 대표적인 AI 비서인 KT Genie 의 "지니야", Apple Siri 의 "Hey, Siri" 와 같은 호출어를 인식하도록 하는 것이다.

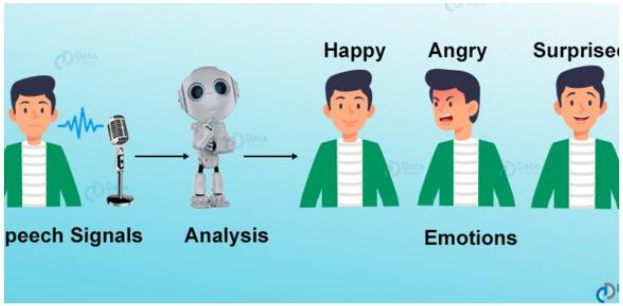
3. Speaker Recognition

사람마다 고유한 음색이 존재한다는 사실을 바탕으로 음성 데이터를 활용하여 사용자 판별 및 사용자 별 맞춤 서비스를 제공할 수 있다.



4. Emotion Recognition

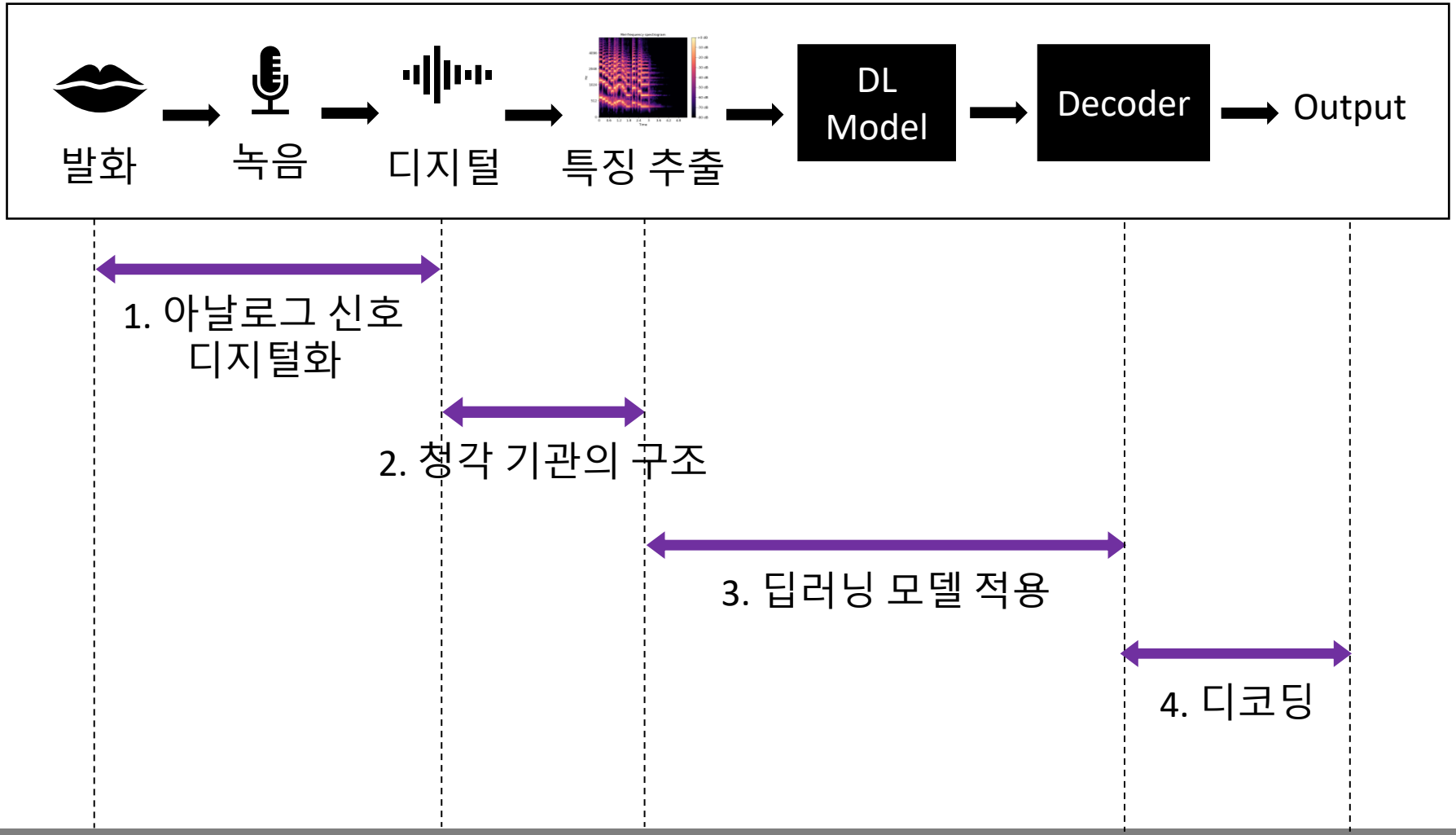
목소리의 변화를 감지하여 사용자의 감정을 파악할 수 있는 방식이다.



목차

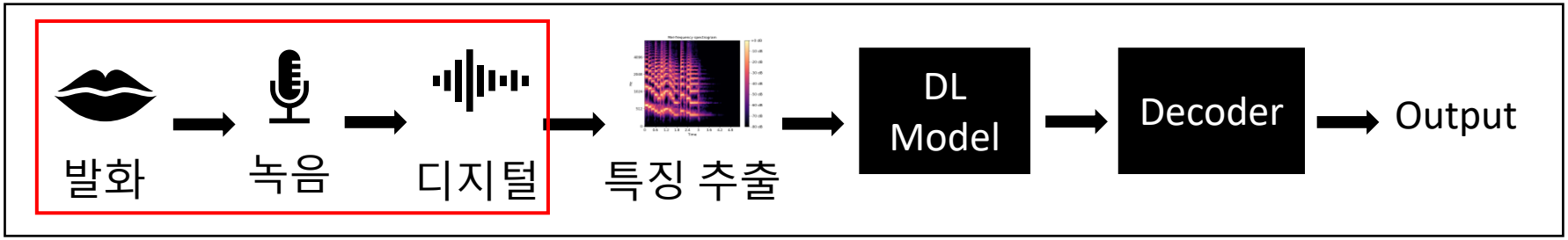
2. End-to-End ASR 딥러닝 모델 원리

- Overview

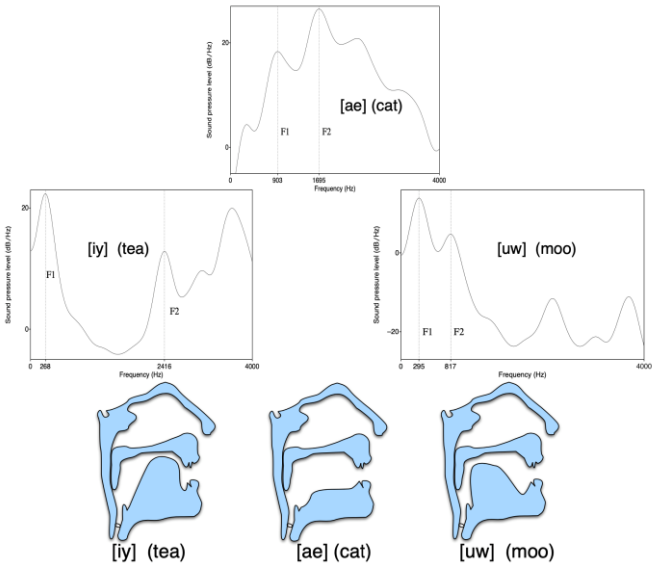


2. End-to-End ASR 딥러닝 모델 원리

- Overview

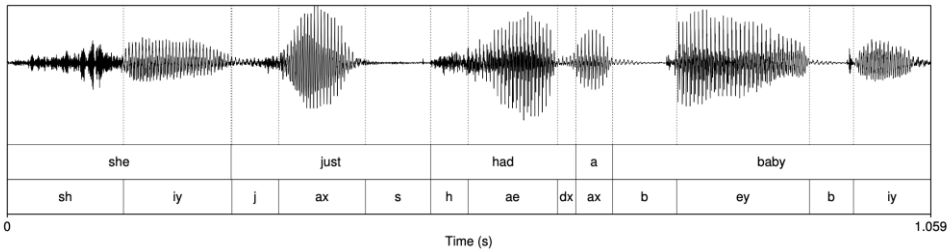


1. 아날로그 신호 디지털화



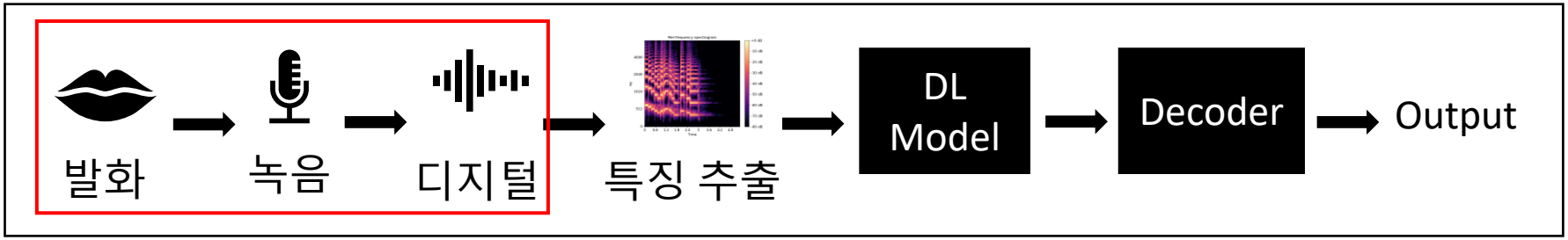
파동(Wave)는 진동하는 신호이며, 대표적인 3 가지 속성으로 진폭(Amplitude), 진동수(Frequency), 위상(Phase)이 있다.

자음, 모음에 따라 파동의 형태가 달라지는 것을 확인할 수 있다.

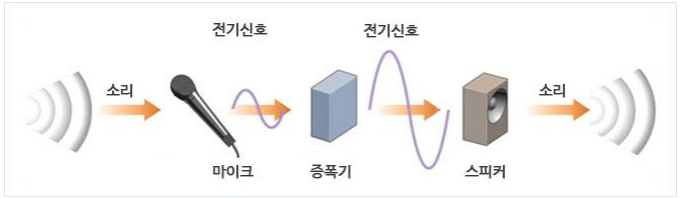


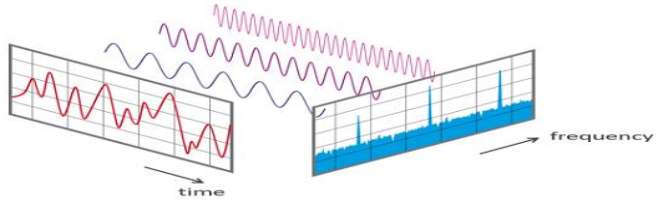
2. End-to-End ASR 딥러닝 모델 원리

- Overview



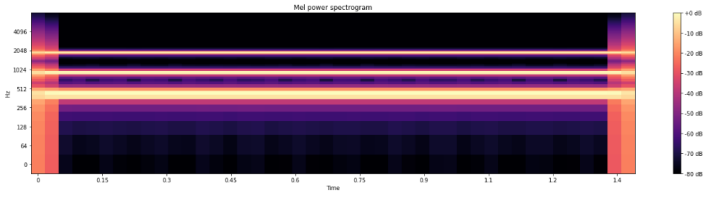
1. 아날로그 신호 디지털화

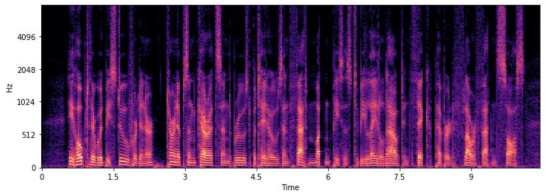




사람 목소리의 떨림이 매질의 진동으로 울려 퍼지면, 마이크의 진동판 떨림을 수치로 저장.

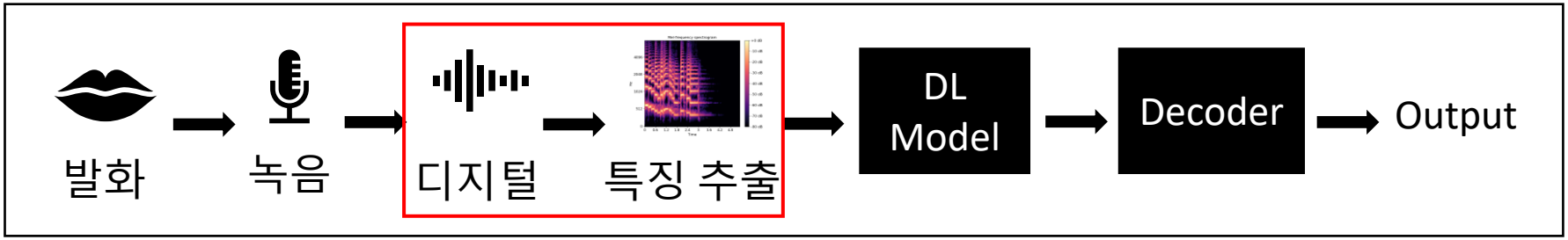
수집된 파동 데이터를 디지털 신호 처리 기법인 푸리에 변환 (Fourier Transform) 을 바탕으로 Spectrum 화 (진동수 파악) 한다.





2. End-to-End ASR 딥러닝 모델 원리

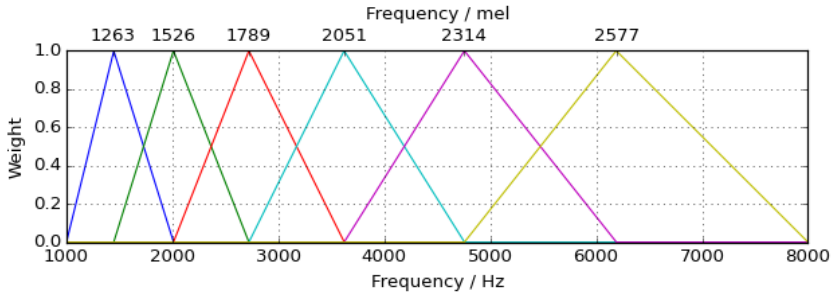
- Overview

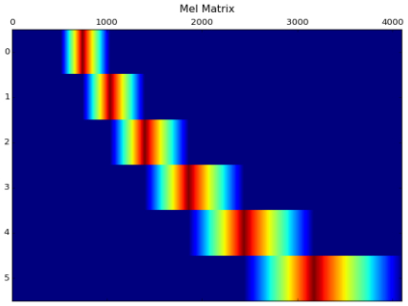


2. 청각 기관의 구조 적용

사람의 청각 기관은 고주파의 변화보다 저주파의 변화에 민감하다.

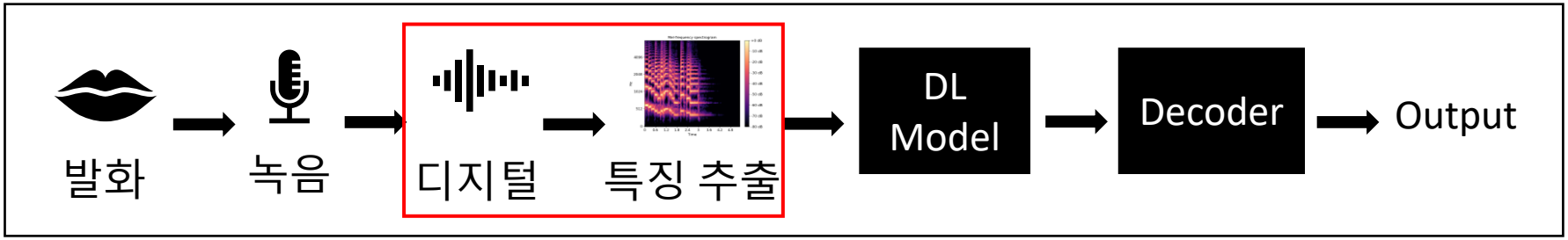
위의 원리를 바탕으로 고주파의 변화보다 저주파의 변화를 더 민감하게 반영하는 Mel Filter Bank 를 Spectrum 에 적용한다.





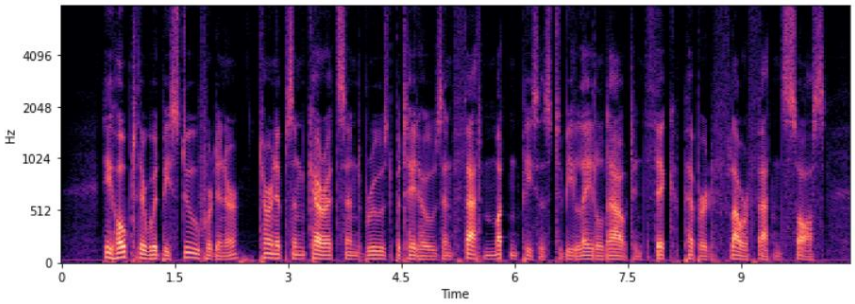
2. End-to-End ASR 딥러닝 모델 원리

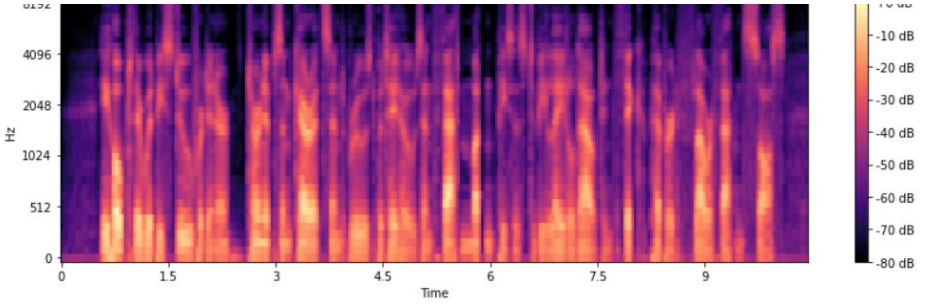
- Overview



2. 청각 기관의 구조 적용

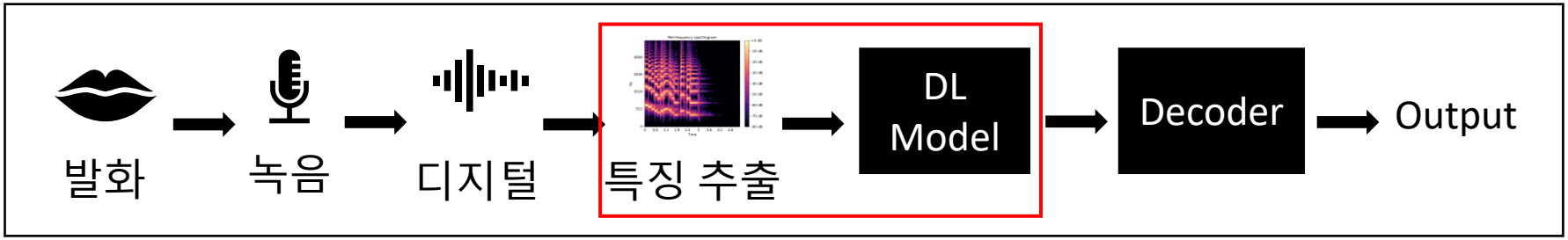
왼쪽의 기본 Spectrum 과 오른쪽의 Mel Filter Bank 를 적용한 Spectrum 을 비교하면 아래 자료와 같이 저음역대의 소리의 영역이 상대적으로 확장된다.





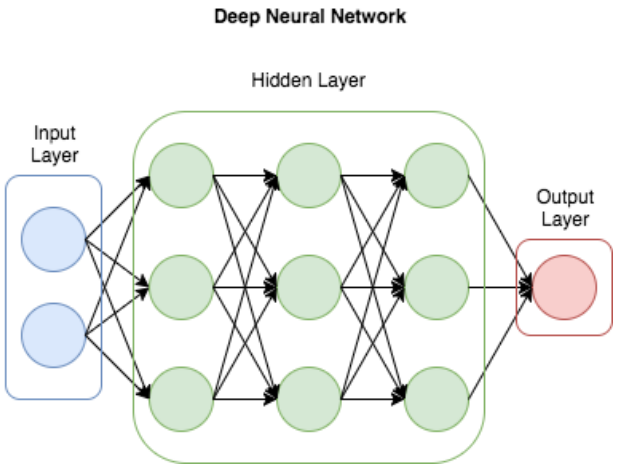
2. End-to-End ASR 딥러닝 모델 원리

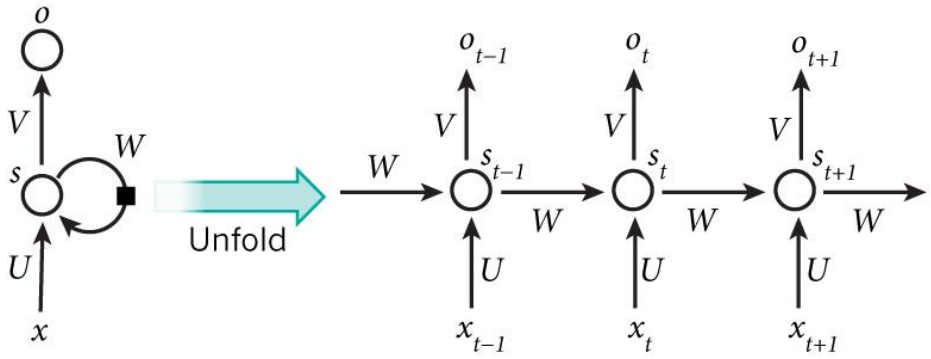
- Overview



3. 딥러닝 모델 적용

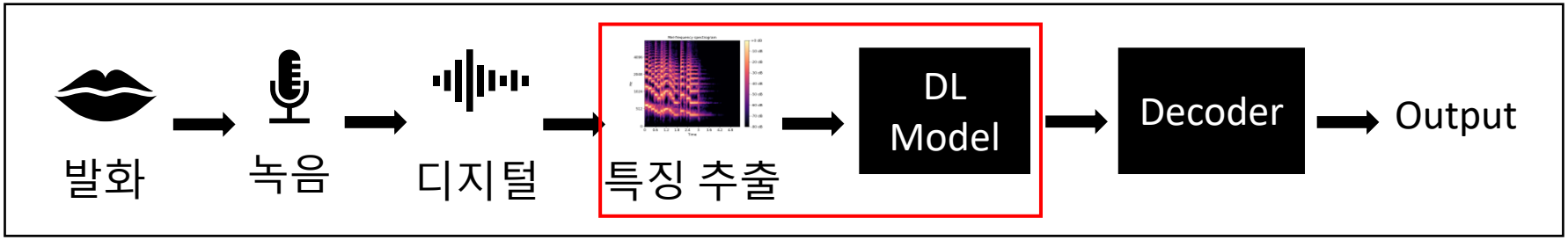
DNN/RNN 등이 적용된 모델을 바탕으로 학습을 진행한다.





2. End-to-End ASR 딥러닝 모델 원리

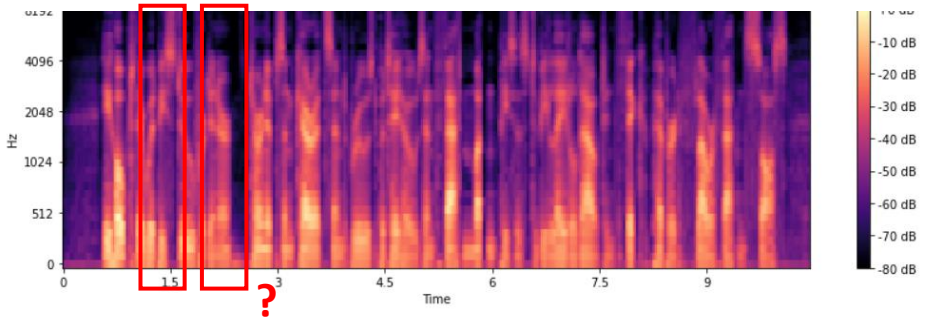
- Overview



3. 딥러닝 모델 적용

학습 중 디지털화된 음성 데이터와 단순 라벨링된 데이터 사이에서 alignment(싱크) 데이터 부재로 인하여 문제점을 만날 수 있다.

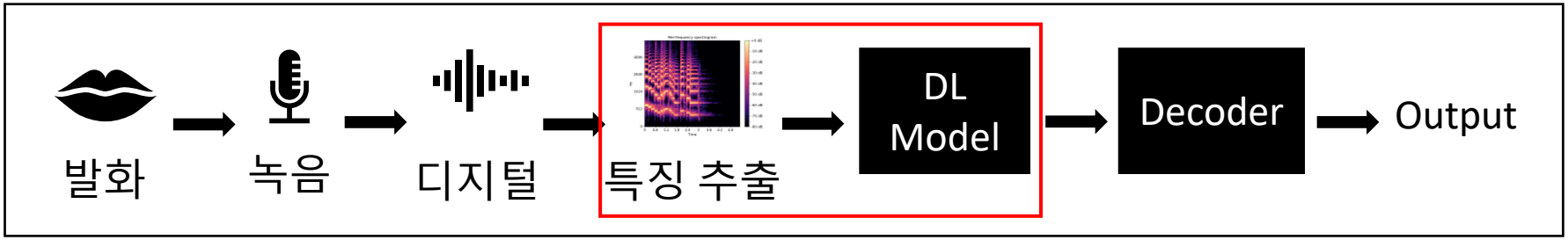
문장 단위에서 해당 단어에 대한 스펙트럼이 어떤 것인가? → **해결책 필요**



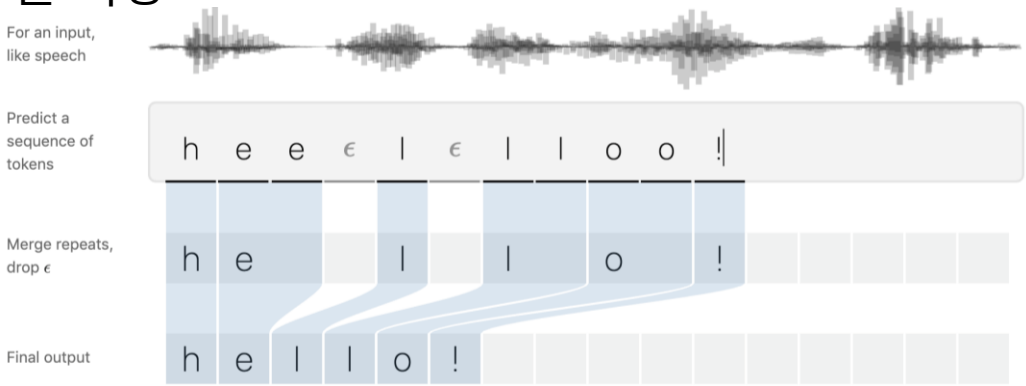
'HE HOPED THERE WOULD BE STEW FOR DINNER TURNIPS AND CARROTS'

2. End-to-End ASR 딥러닝 모델 원리

- Overview



3. 딥러닝 모델 적용

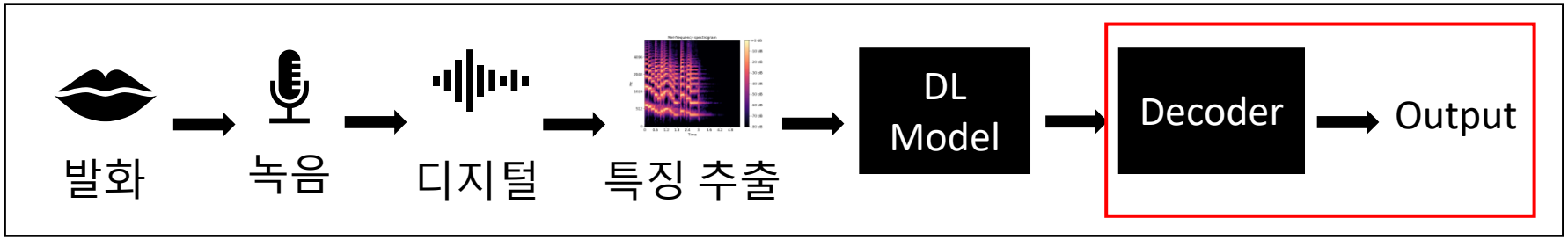


이에 대한 방안으로 CTC (Connectionist Temporal Classification) 가 있다.

Input과 결과 데이터 사이에서 alignment 정보가 없어도 학습이 가능하다.

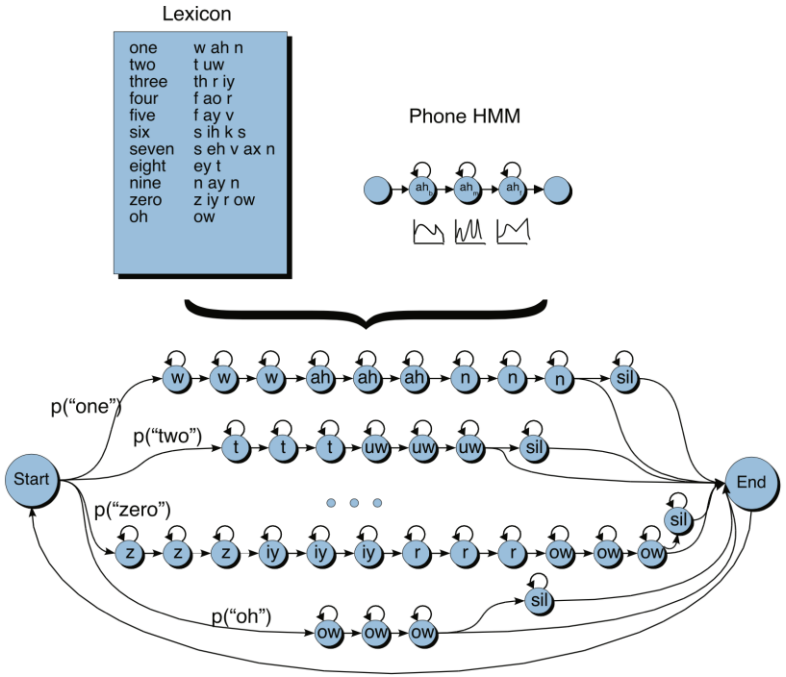
2. End-to-End ASR 딥러닝 모델 원리

- Overview

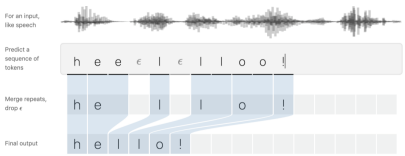


4. 디코딩 적용

- Acoustic + Language Model
  - 어휘사전 (lexicon)

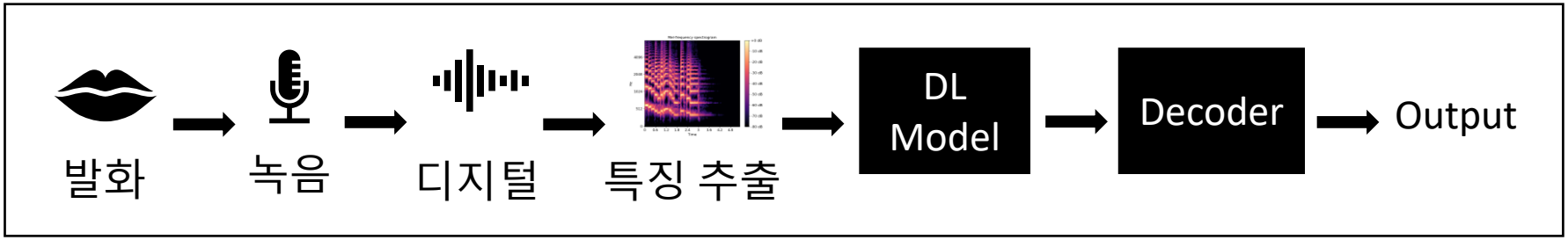


- Sequence Model
  - CTC



2. End-to-End ASR 딥러닝 모델 원리

- Overview



5. 모든 과정이 끝나고..

후처리(Post-Processing) 기법을 통해 더 읽기 좋은 텍스트로 만든다.

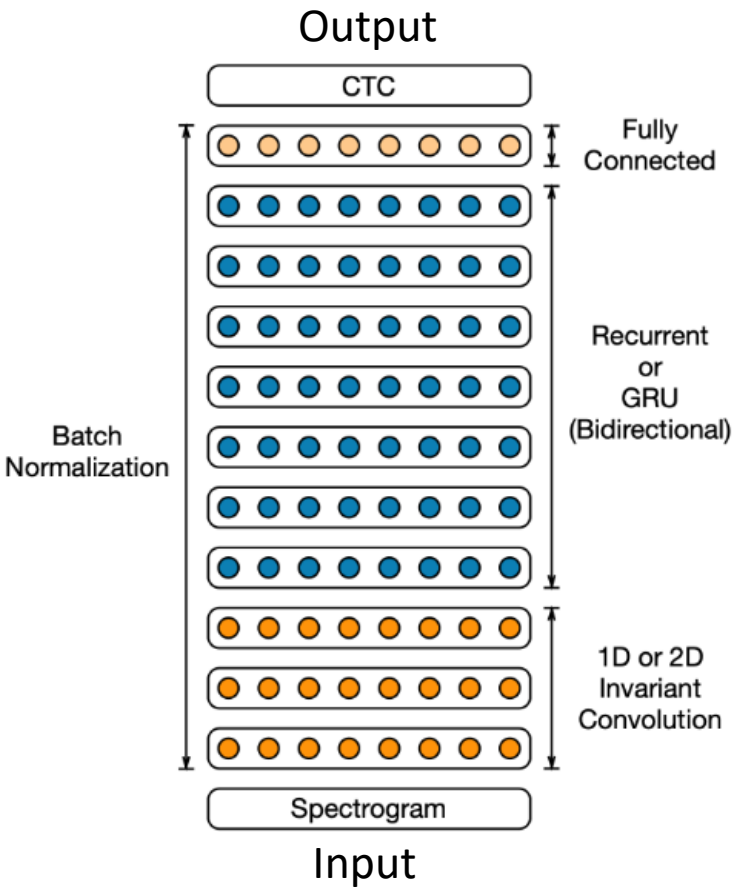
위 모든 과정이 끝나면 Output 결과를 내놓게 된다.

Evaluation 에는 Word Error Rate(WER), Sentence Error Rate(SER) 등이 있다.

목차

3. Deep Speech 2 모델

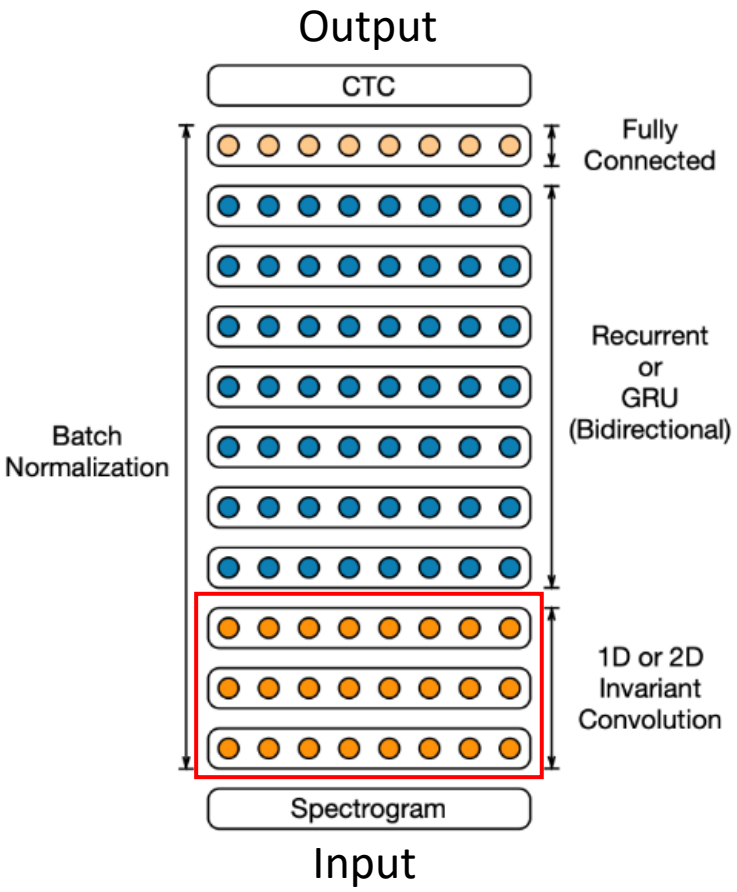
1. Deep Speech 2 모델 구조



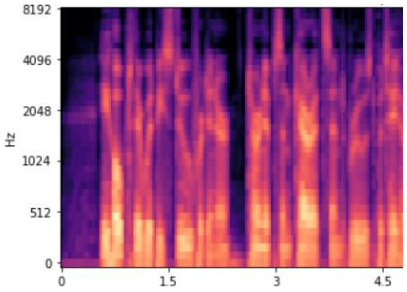
- **CNN Layer** x 3
  - 특징 추출
  - 1D or 2D Convolution
- **Bi-directional RNN Layer** x 7
  - Vanilla RNN 개선 모델
  - 순차적 데이터 학습
- **Fully Connected Layer** x 1
- **CTC** x 1
- **Batch Normalization**
  - On every layer
  - SortaGrad

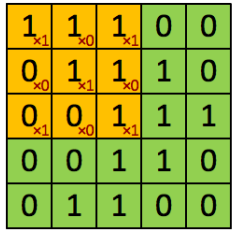
3. Deep Speech 2 모델

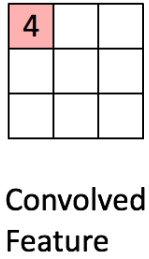
1. Deep Speech 2 모델 구조



- **CNN Layer** x 3
  - 특징 추출
  - 1D / 2D Convolution
- **원리**
  - 보통 25ms 단위의 프레임으로 구성
  - 음성 데이터를 바탕으로 CNN 학습



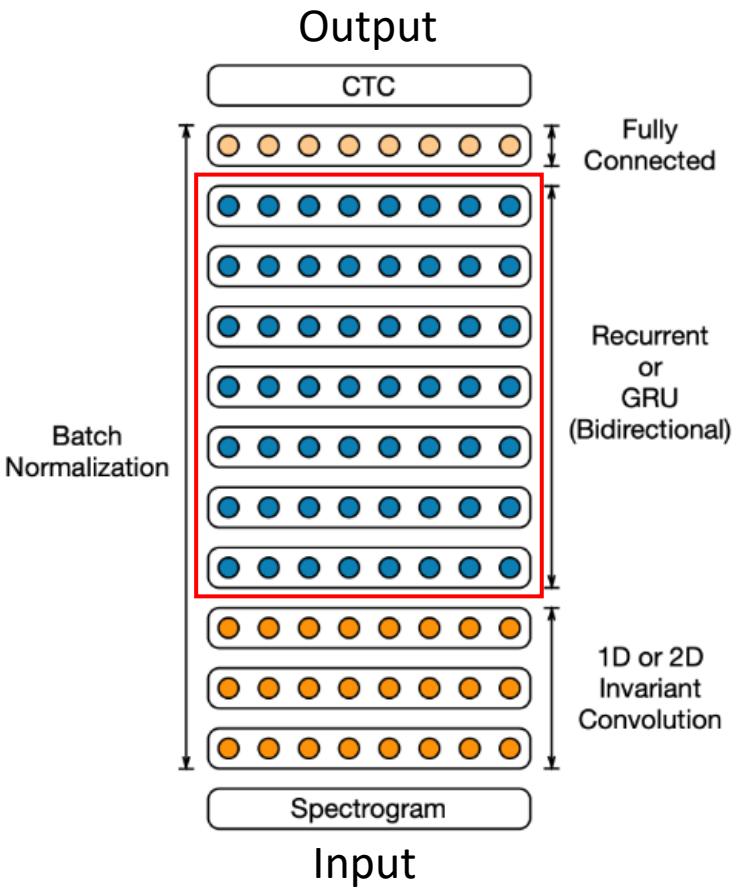




$$f(x) = \min\{\max(x, 0), 20\}$$

3. Deep Speech 2 모델

1. Deep Speech 2 모델 구조



- **Bi-directional RNN Layer** x 7
  - 일반 RNN 개선 모델
  - 순차적 데이터 학습에 적합

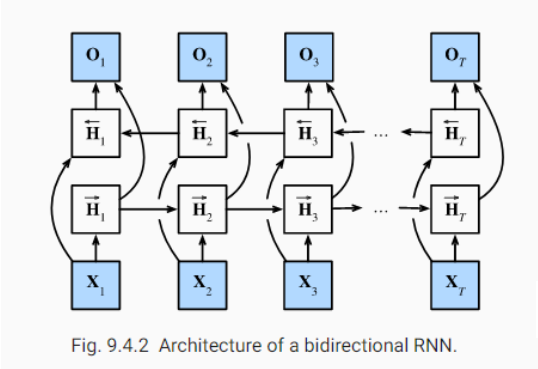


Fig. 9.4.2 Architecture of a bidirectional RNN.

$$\overrightarrow{h_t^l} = g\left(h_t^{l-1}, \overrightarrow{h_{t-1}^l}\right)$$

$$\overleftarrow{h_t^l} = g\left(h_t^{l-1}, \overleftarrow{h_{t+1}^l}\right)$$

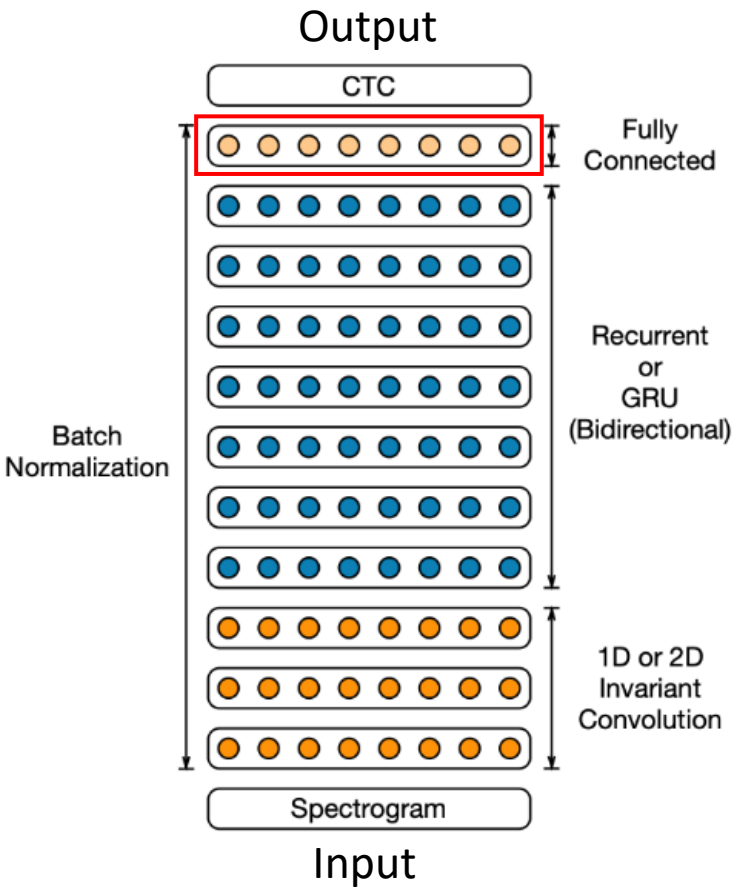
$$\overrightarrow{h_t^l} = f\left(W^l h_t^{l-1} + \overrightarrow{U^l h_{t-1}^l} + b^l\right)$$

$$f(x) = \min\{\max(x, 0), 20\}$$

- **장점**
  - 전후 데이터 모두를 반영할 수 있음
    - i) I am ____.
    - ii) I am ____ hungry.
    - iii) I am ____ hungry, so now I could eat anything.
- **단점**
  - 시간 복잡도가 올라감

3. Deep Speech 2 모델

1. Deep Speech 2 모델 구조



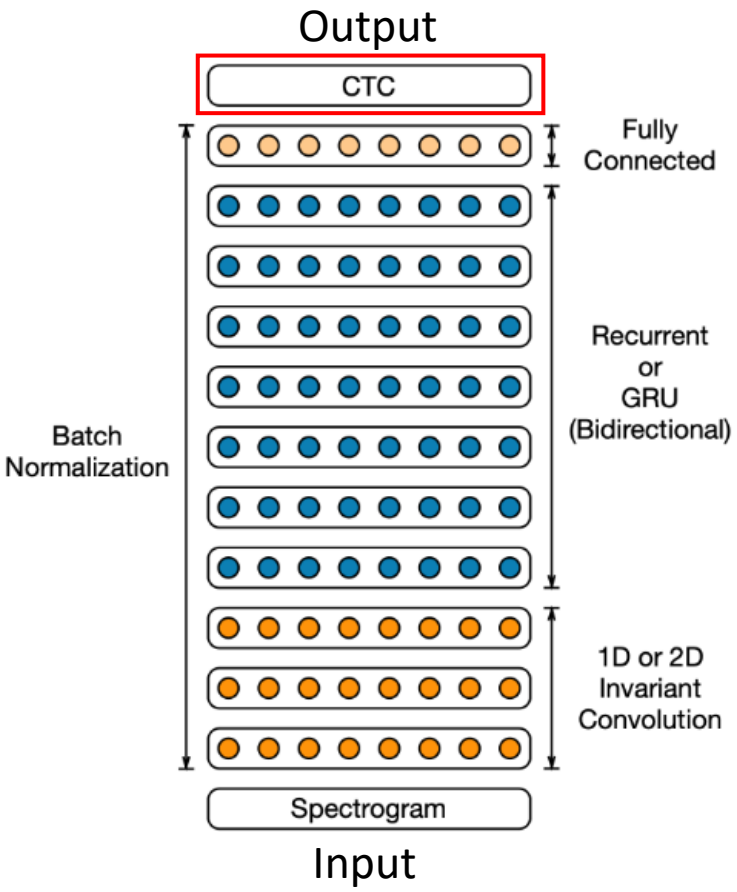
- **Fully Connected Layer** x 1
- **구조**
  - 이전 학습 데이터들에 대하여 1차원 신경망에 적용
  - Softmax 함수 적용

$$h_t^l = f\left(W^l h_t^{l-1} + b^l\right)$$

$$p\left(l_t = k|x\right) = \frac{\exp\left(w_j^L h_t^{L-1}\right)}{\sum_j \exp\left(w_j^L h_t^{L-1}\right)}$$

3. Deep Speech 2 모델

1. Deep Speech 2 모델 구조

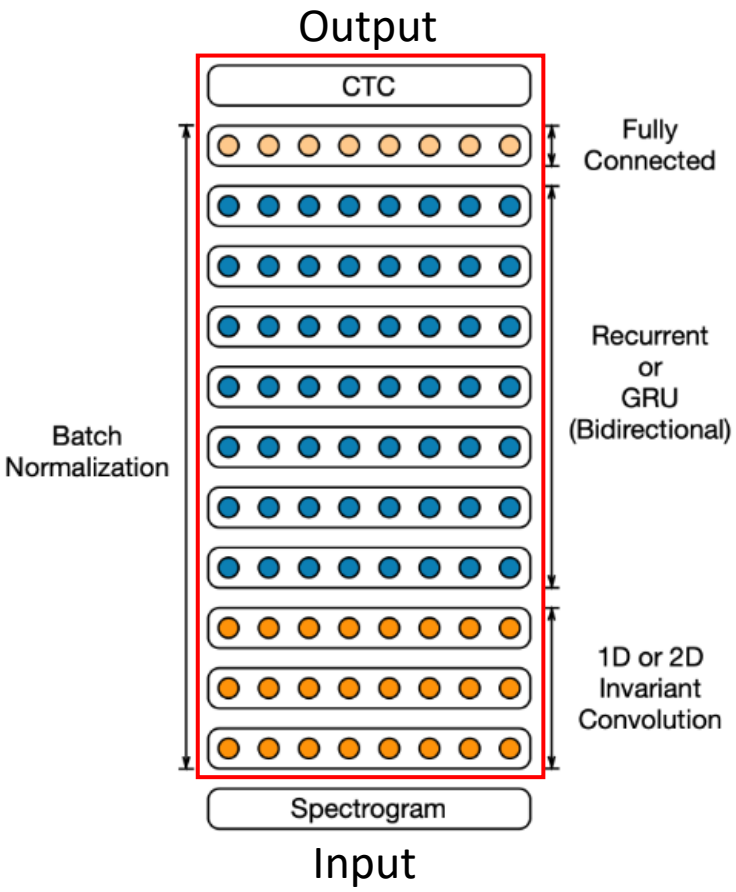


- **CTC** x 1
- **구조**
  - CTC loss 계산을 통하여 역전파(backpropagation)을 바탕으로 모델을 학습함.

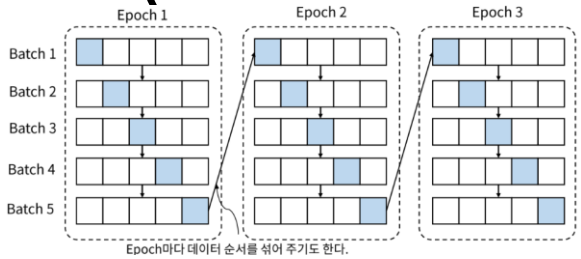
$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial h_1} \times \frac{\partial h_1}{\partial z_1} \times \frac{\partial z_1}{\partial w_1} \qquad w_1^+ = w_1 - \alpha \frac{\partial E_{total}}{\partial w_1}$$

3. Deep Speech 2 모델

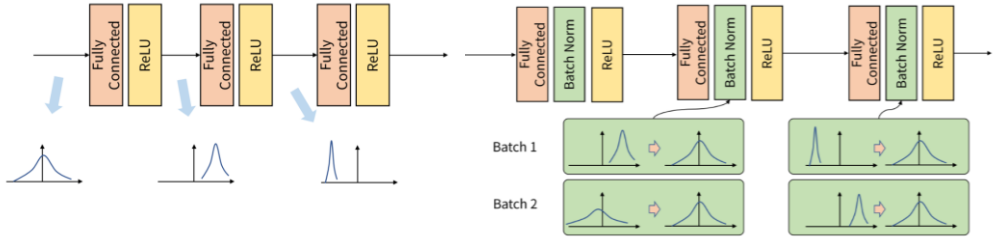
1. Deep Speech 2 모델 구조



- **SGD (Stochastic Gradient Descent)**

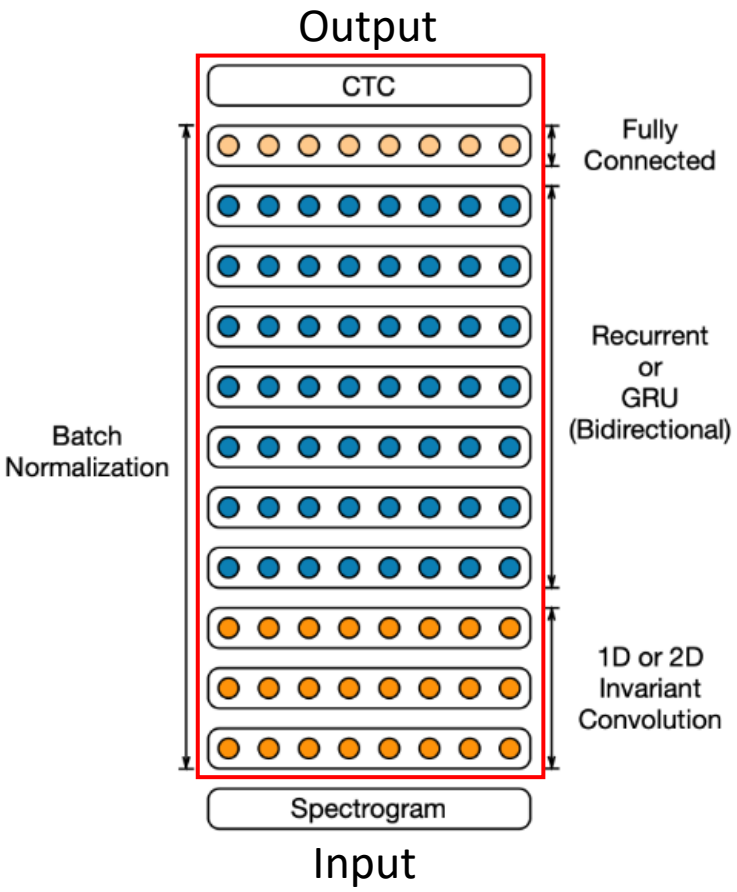


- **Batch Normalization**



3. Deep Speech 2 모델

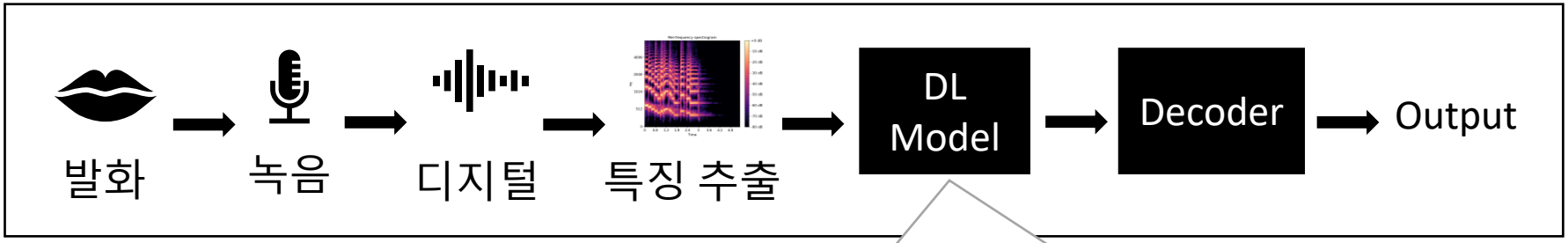
1. Deep Speech 2 모델 구조

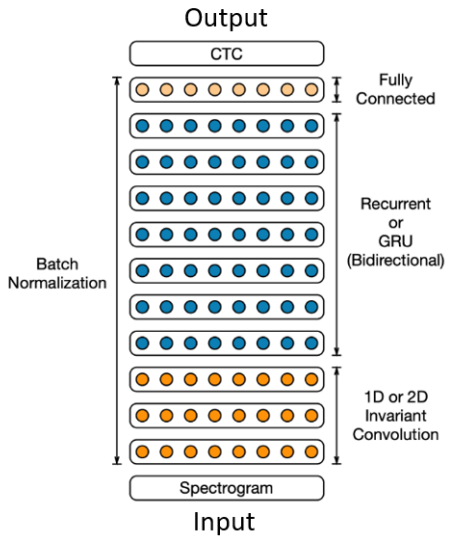


- **SortaGrad (in CTC loss)**
- 다양한 데이터를 무작위로 학습시키면서 그래디언트 길이가 불균형해짐.
- 이를 개선하기 위해 음성 데이터 크기를 오름차순으로 정렬한 후 CTC 학습을 진행하는 방식을 채택함.

3. Deep Speech 2 모델

- Summary





- **CNN Layer** x 3
  - 특징 추출
  - 1D or 2D Convolution
- **Bidirectional RNN Layer** x 7
  - Vanilla RNN 개선 모델
  - 순차적 데이터 학습
- **Fully Connected Layer** x 1
- **CTC** x 1
- **Batch Normalization**
  - On every layer
  - SortaGrad

Thank you!

MILab Undergraduate student, Kim Taehyeon

2023. 04. 21

Reference

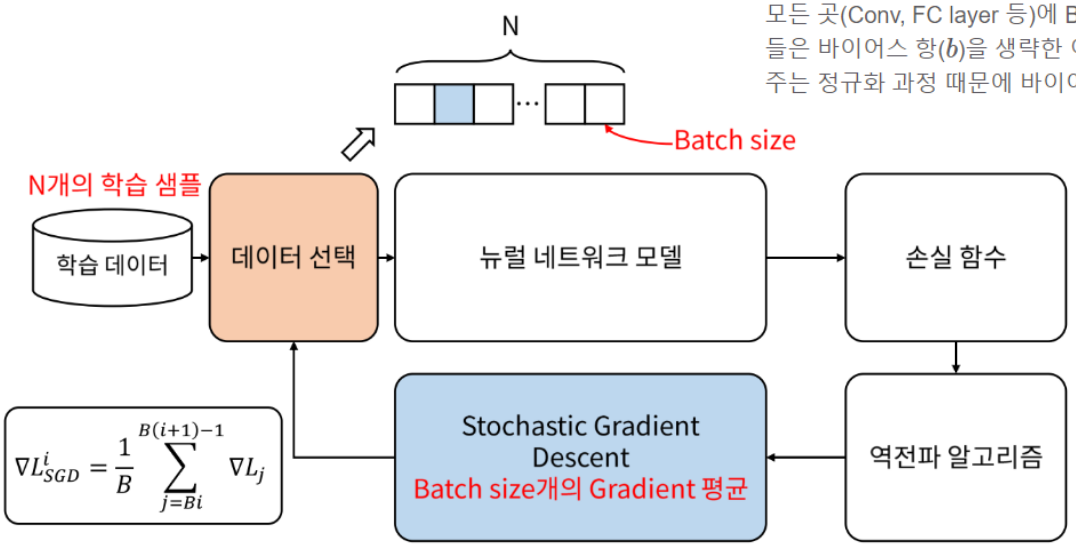
- https://www.mdpi.com/2079-9292/9/7/1157
- https://learn.microsoft.com/ko-kr/azure/cognitive-services/speech-service/speaker-recognition-overview
- https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/
- https://siggigue.github.io/pyfilterbank/melbank.html
- https://ratsgo.github.io/speechbook/
- https://www.nti-audio.com/ko/
- https://hyen4110.tistory.com/29
- https://wikidocs.net/37406
- https://gaussian37.github.io/dl-concept-batchnorm/

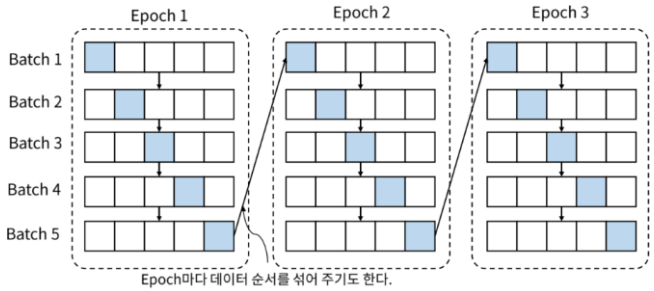
풀이과정

$$\frac{W-F+2P}{S}+1 = \frac{7-3+2\cdot 0}{1}+1$$

$$= 4+1 = 5$$

`Deep Speech2` 저자들은 선형변환(linear transformation)과 활성함수 f(clipped ReLU)가 연이어 나타나는 모든 곳(Conv, FC layer 등)에 Batch Normalization을 추가했습니다 : $f(Wh+b) \rightarrow f(\mathcal{B}(\mathcal{W}h))$. 저자들은 바이어스 항($b$)을 생략한 이유로 Batch Normalization을 수행하면 이동평균(moving average)을 빼주는 정규화 과정 때문에 바이어스 효과가 상쇄되기 때문이라고 설명합니다.





$$\mu_{batch} = \frac{1}{B}\sum_i x_i$$

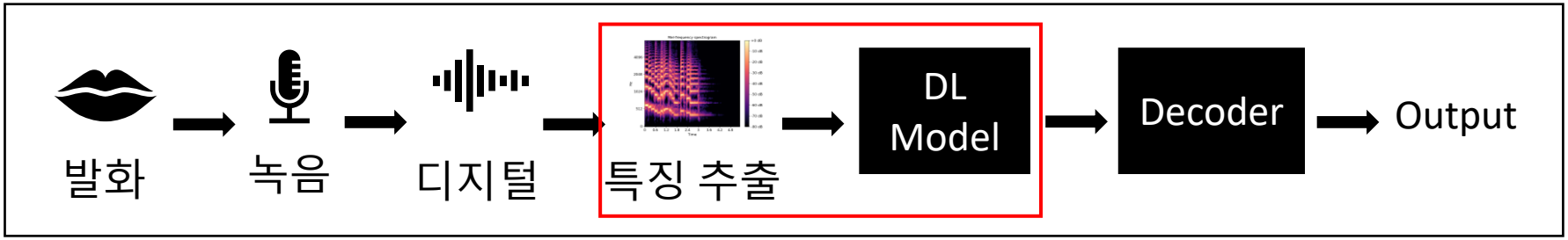
$$BN(X) = \gamma\left(\frac{X-\mu_{batch}}{\sigma_{batch}}\right)+\beta$$

$$\sigma^2_{batch} = \frac{1}{B}\sum_i (x_i-\mu_{batch})^2$$

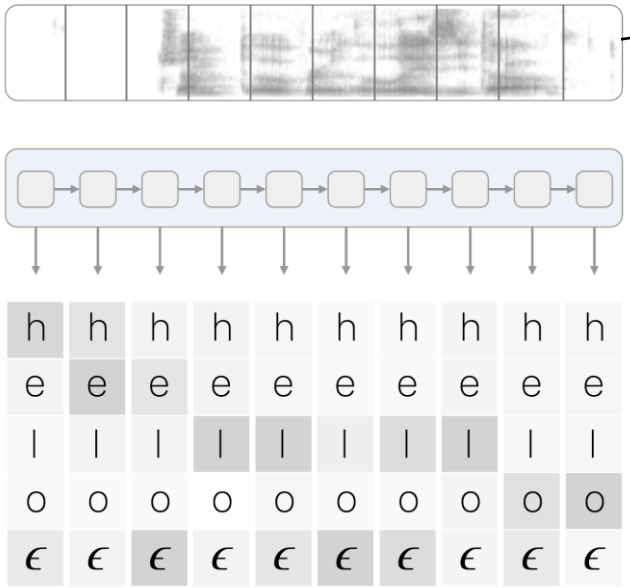
$$\overrightarrow{h^l_t} = f\left(\mathcal{B}\left(\mathcal{W}^l h^{l-1}_t\right) + \overrightarrow{\mathcal{U}^l h^l_{t-1}} + b^l\right)$$

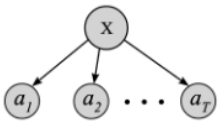
2. End-to-End ASR 딥러닝 모델 원리

- Overview



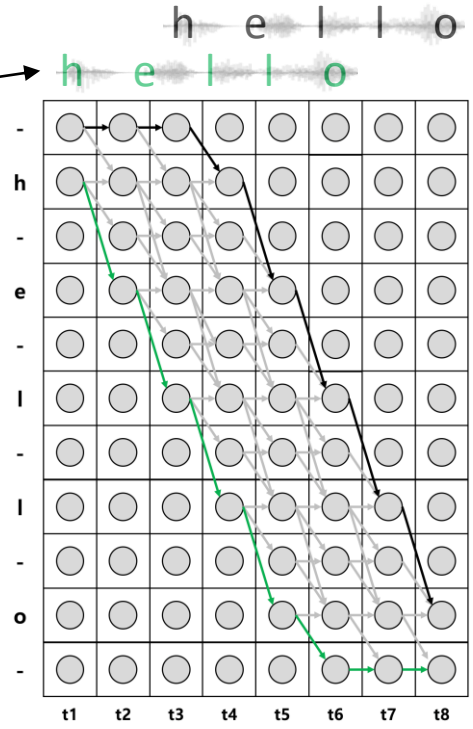
3. 딥러닝 모델 적용





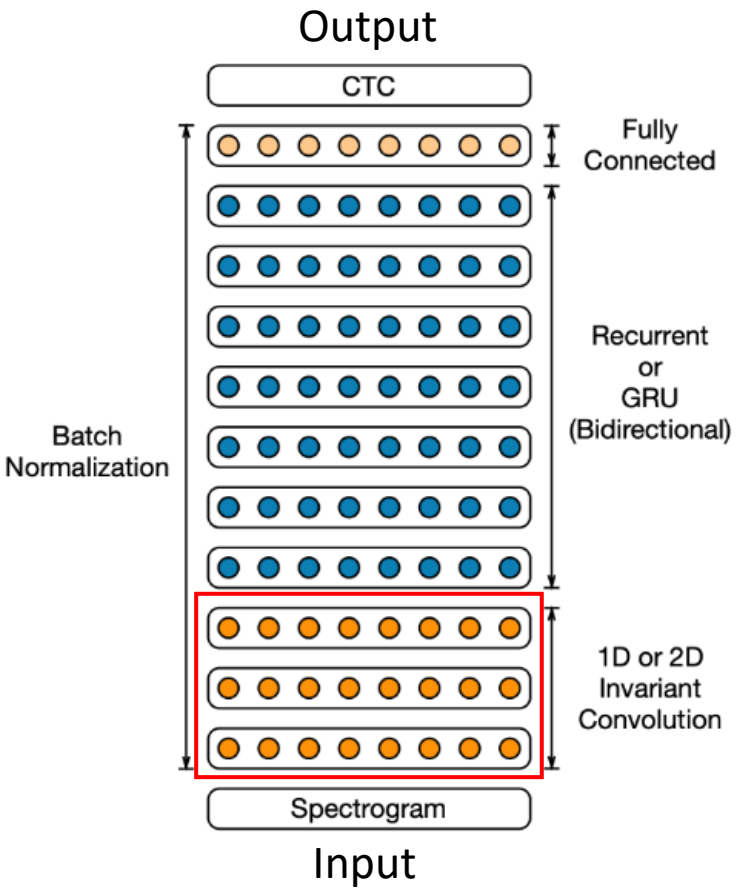
$$p(\pi|\mathbf{x}) = \prod_{t=1}^{T} y_{\pi_t}^t$$

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{x})$$

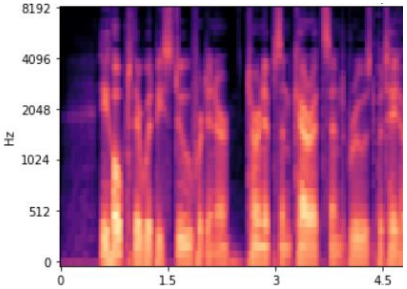


3. Deep Speech 2 모델

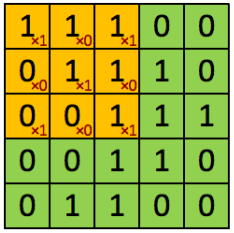
1. Deep Speech 2 모델 구조



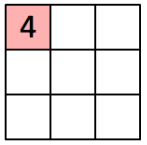
- **CNN Layer** x 3
  - 특징 추출
  - 1D or 2D Convolution (2D is better)
- **원리**
  - 보통 25ms 단위의 프레임으로 구성
  - Mel Filter Bank 를 통해 (count_of_frames) x (n_mels) 의 차원 데이터를 바탕으로 CNN 학습



Dimension
= (4.5sec / 25msec, n_mels)
= (180, 40)



Image



Convolved Feature

$$f(x) = \min\{\max(x, 0), 20\}$$