



NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS : Tacotron2

고급 심화 세션 강효은


Motivation




Motivation

YouTube^{KR}

브루노마스 ai hype boy



Hype Boy - Bruno Mars (Original by Newjeans) (AI COVER)

WhoAml AiCover

구독자 1.7만명

구독

6.7만

공유

오프라인 저장

클립

조회수 220만회 4개월 전

Credit to original vocal - g1nger

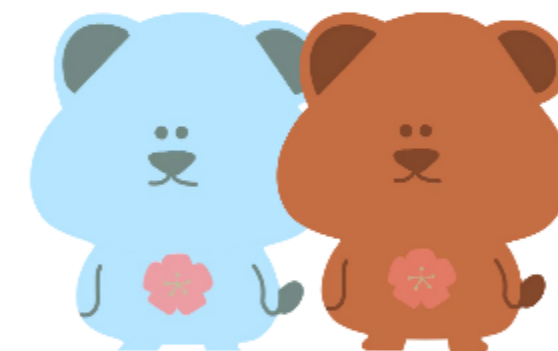
NewJeans 뉴진스 - Hype Boy 하입보이 (cover b...

Don't forget to like and subscribe to the original vocal's owner. 🙏🙏🙏 ...더보기

EWHA

EURON

Background

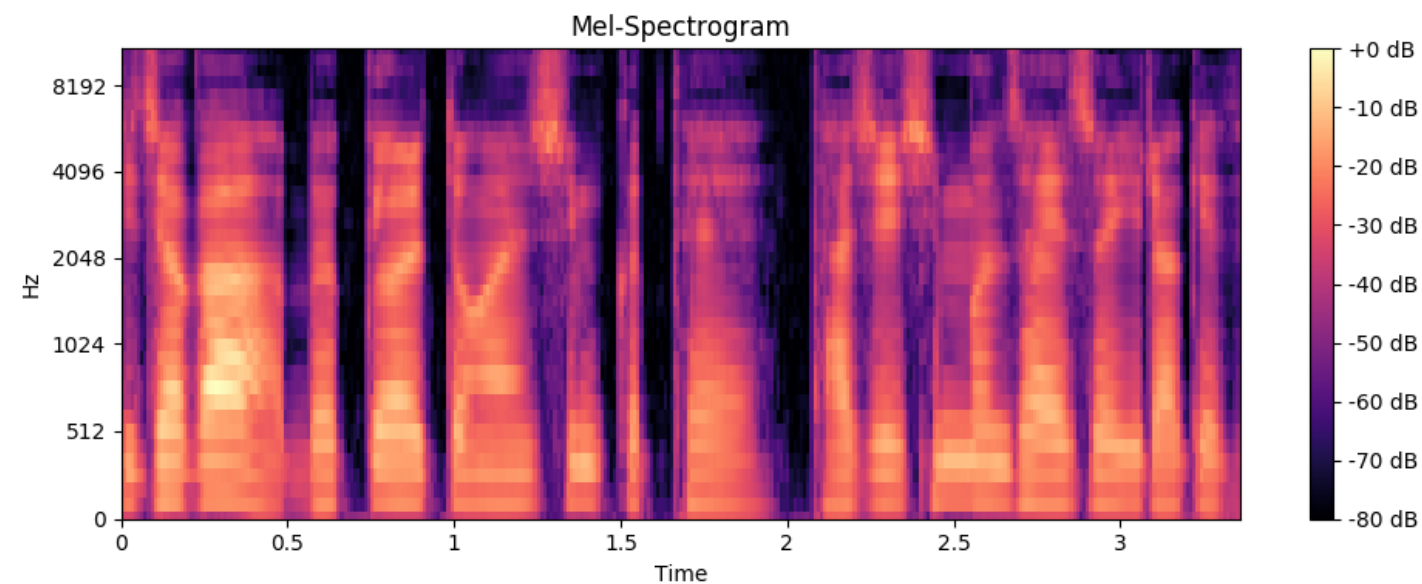


Background

* TTS (Text-to-Speech)

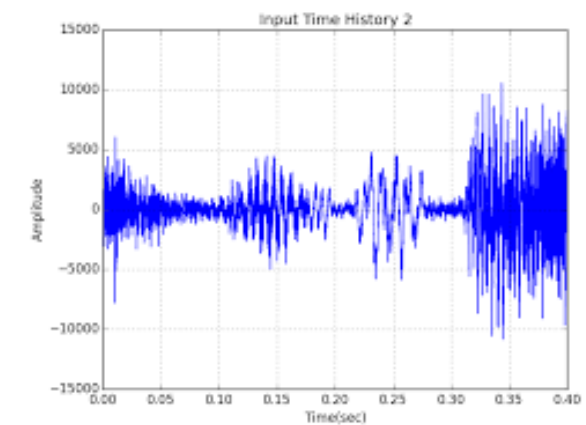
TEXT

text2mel



Mel-Spectrogram

mel2wav



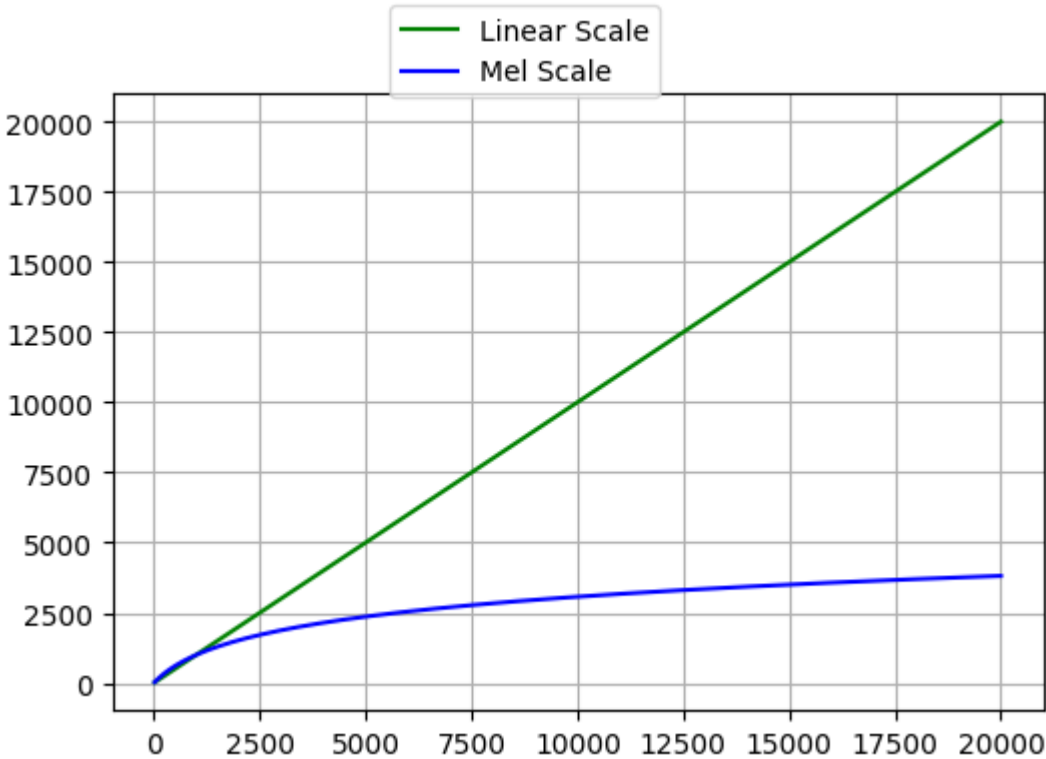
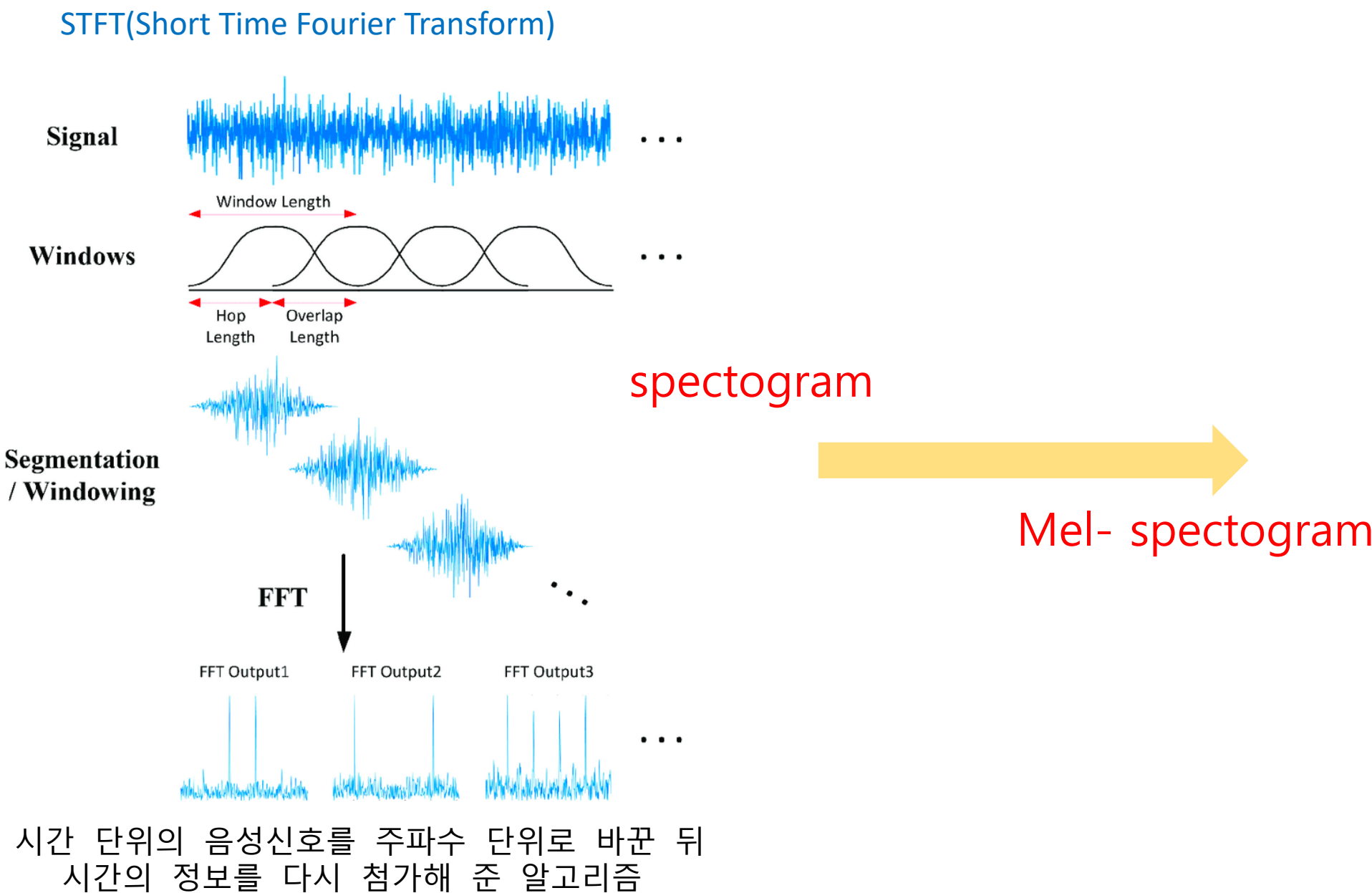
SPEECH

Background

https://www.researchgate.net/publication/346243843_Area-Efficient_Short-Time_Fourier_Transform_Processor_for_Time-Frequency_Analysis_of_Non-Stationary_Signals

* Mel-Spectrogram?

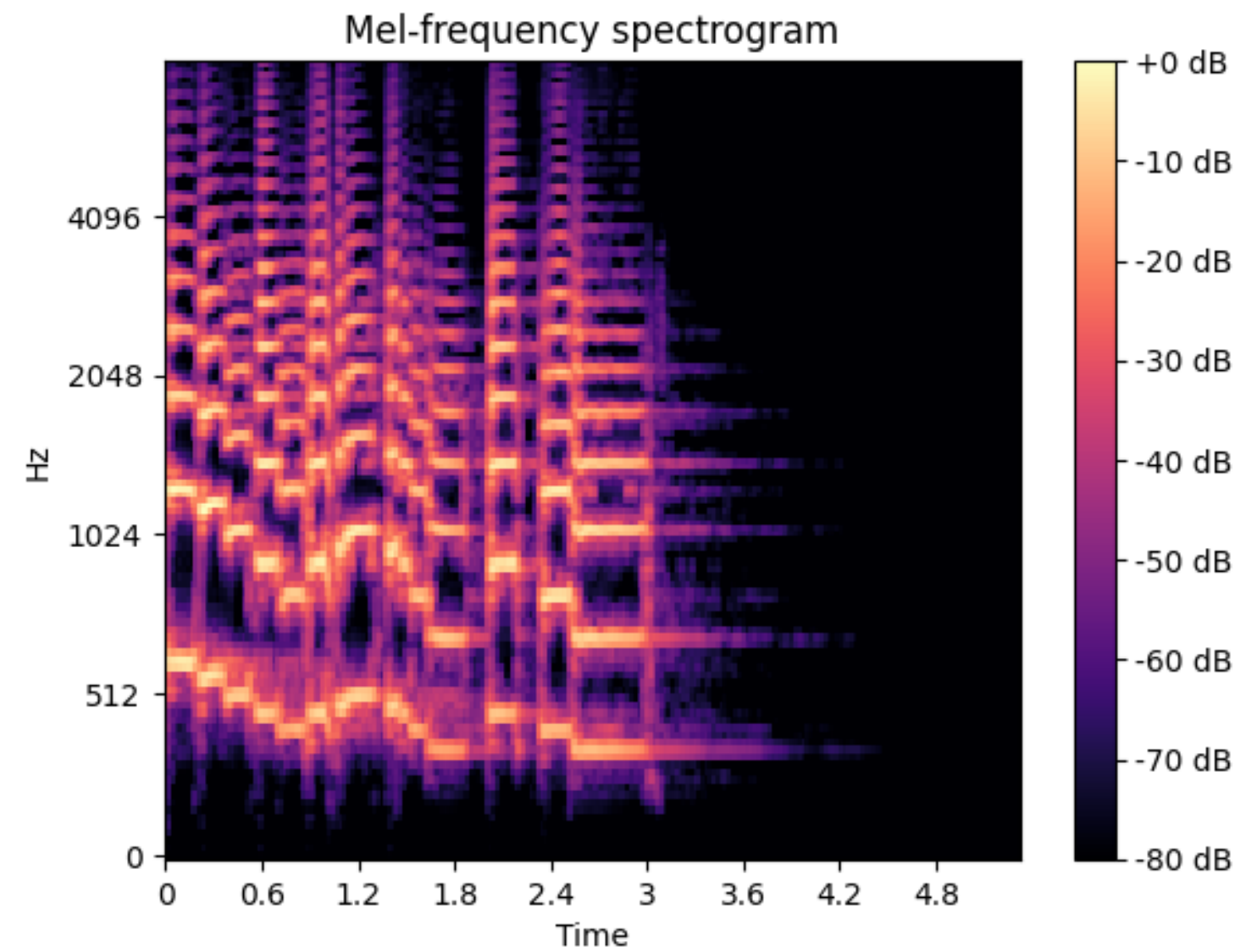
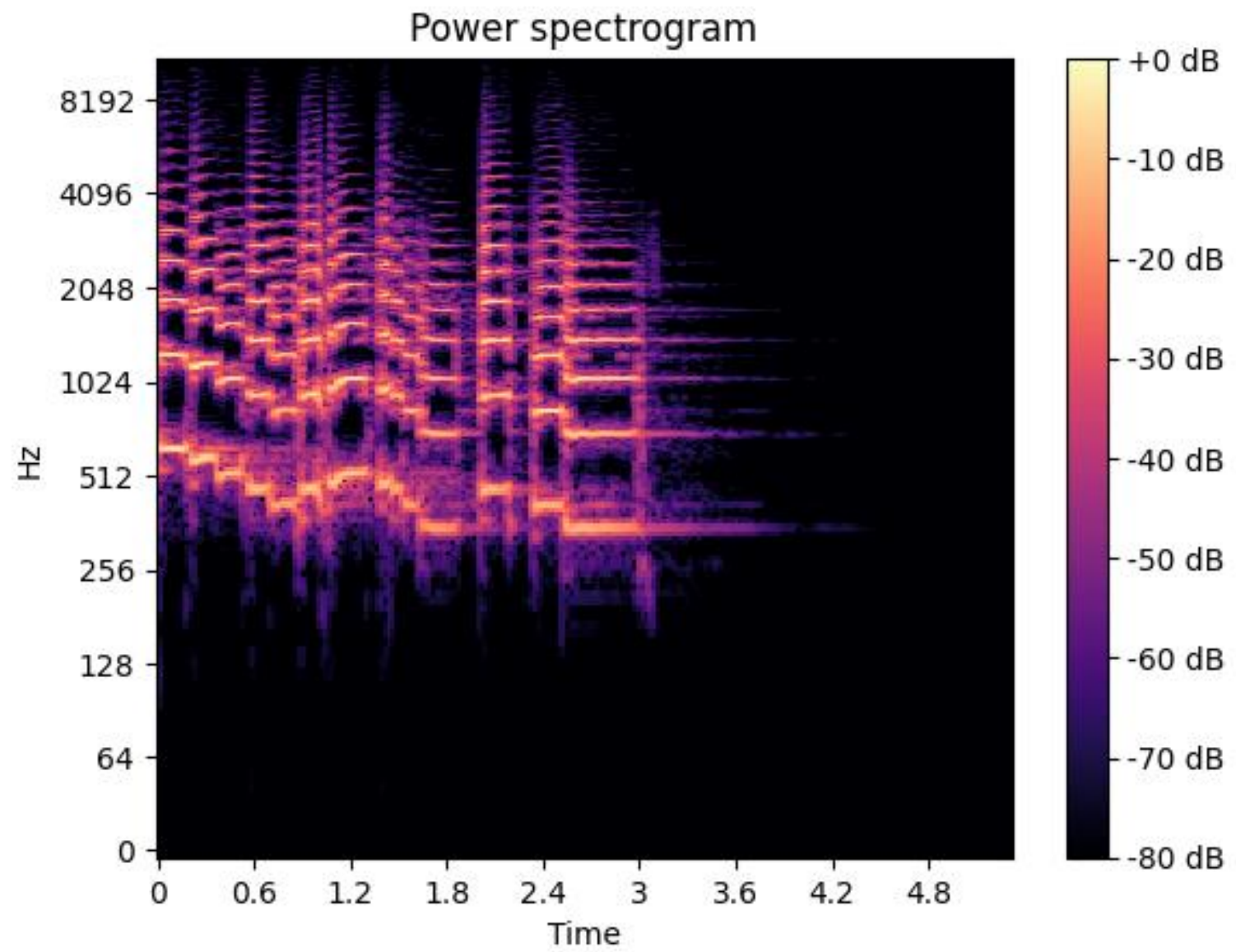
- : 음성의 특징 추출에 사용하는 대표적인 방법 중 하나
- : time-domain의 원본 음성을 frequency-domain으로 바꾸어주는 Fourier Transform(푸리에 변환)에 기초



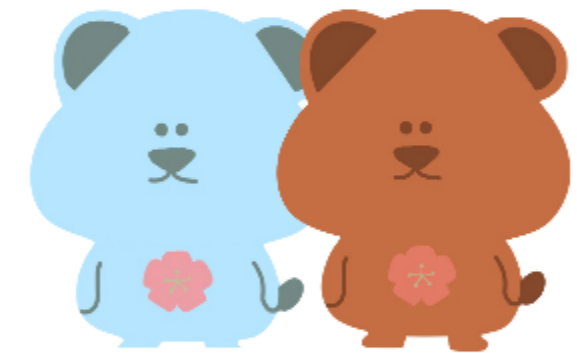
$$m = 2595 \log \left(1 + \frac{f}{700} \right)$$

인간은 음성을 선형적으로 인식하지 않는다.
낮은 소리의 차이에 대해서는 더 예민하게, 높은
소리의 차이에 대해서는 더 둔감하게 듣는 경향 존재

Background



Abstract



Abstract

1. Attention 기반 Seq2Seq의 TTS 모델을 제시
2. <문장, 음성> 쌍으로 이뤄진 데이터만으로 별도 작업없이 학습 가능한 end-to-end 모델
3. MOS (음성 합성 품질 테스트)에서 높은 점수 획득

Introduction



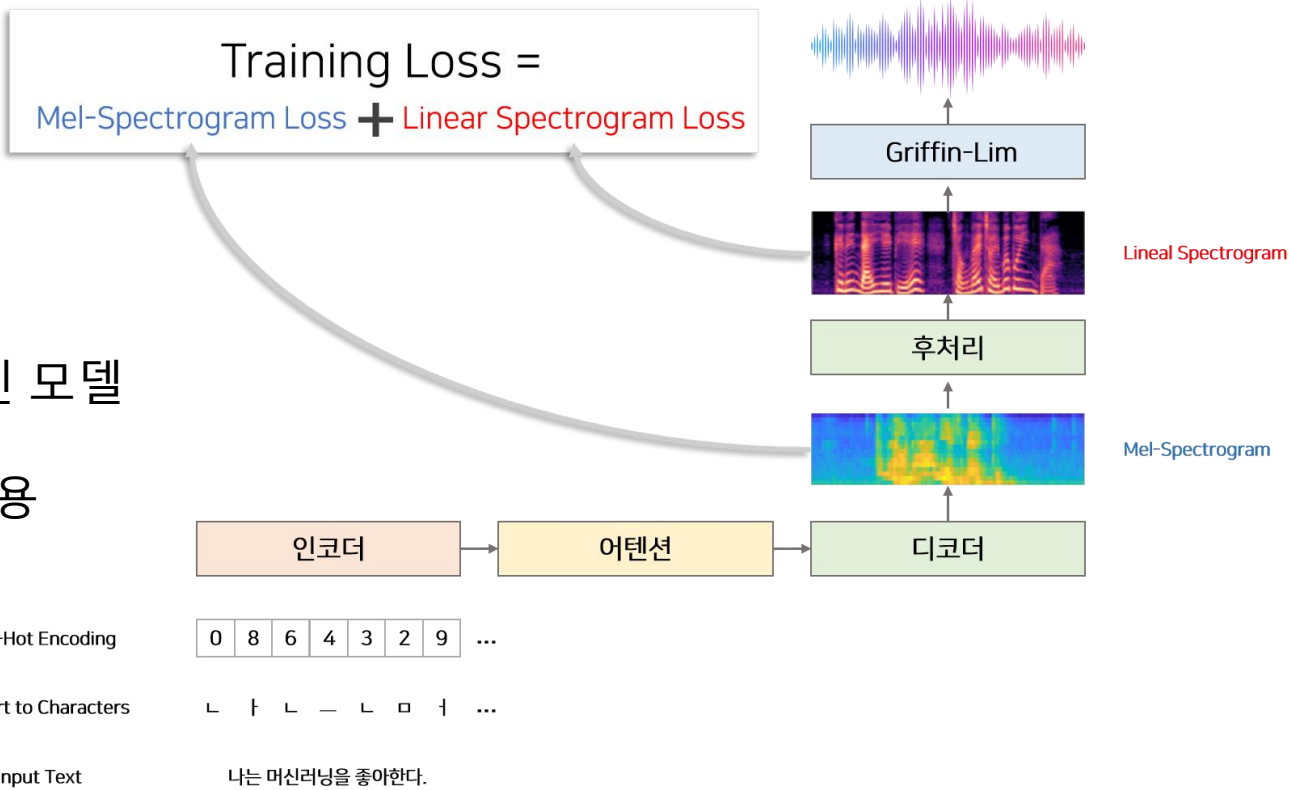
Introduction

<https://joungeekim.github.io/2020/09/25/paper-review/>

Tacotron 2

Tacotron 1

- : Tacotron 1을 계승하되, 모델 성능을 높이기 위해 몇몇 실험을 더 거친 모델
- : Tacotron 1에서는 vocoder로 griffin-Lim 이용



waveNet

- : vocoder (mel2wav)

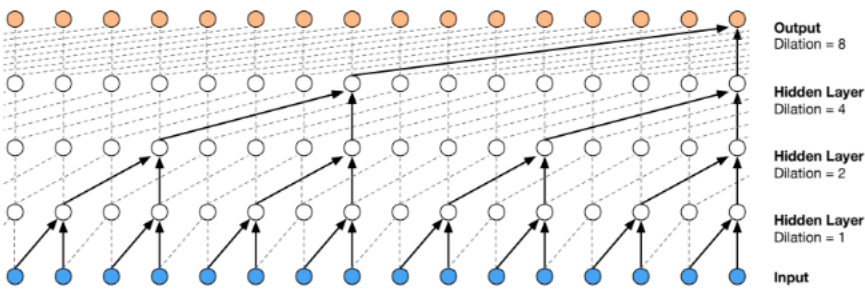
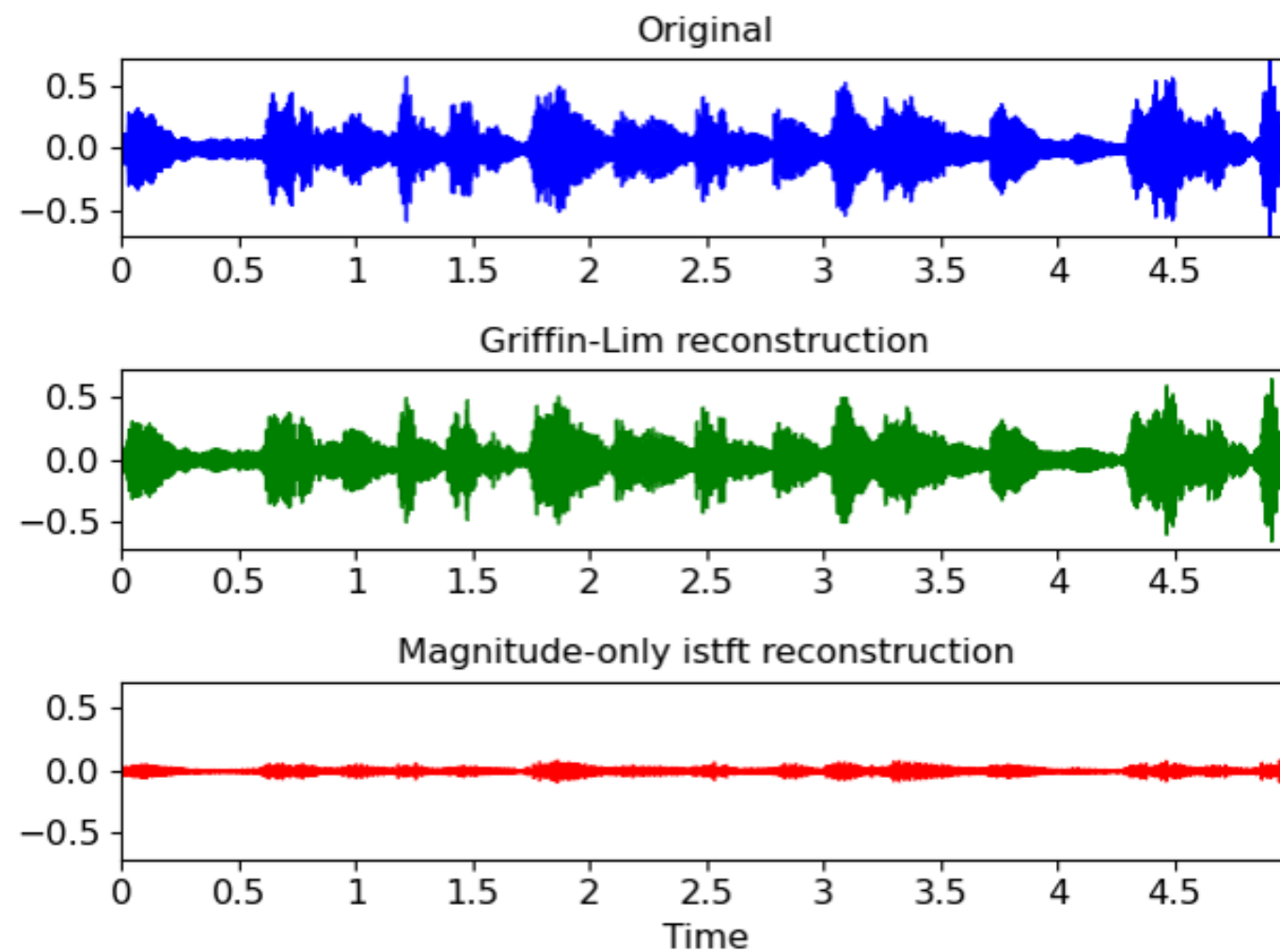


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

Introduction

<https://velog.io/@tobigsvoice1516/3%EC%A3%BC%EC%B0%A8-%EC%9D%8C%EC%84%B1%ED%95%A9%EC%84%B15>

+) griffin-Lim ?



Mel-spectrogram으로 계산된 STFT의 magnitude 값을 통해
원본 음성을 예측하는 rule-based 알고리즘

STFT phase 정보를 구하기 위해 임의의 값으로 두고,
예측된 음성의 STFT magnitude 값과 Mel-spectrogram으로
계산된 STFT magnitude 값의 MSE가 최소가 되도록 반복
하여 원본 음성 찾아감.

이러한 결정적(deterministic) 알고리즘들은
실험에 의해 인위적인 소리를 만들어 낸다는 비판

+ 오디오 합성 품질이 좋지 않음

Model



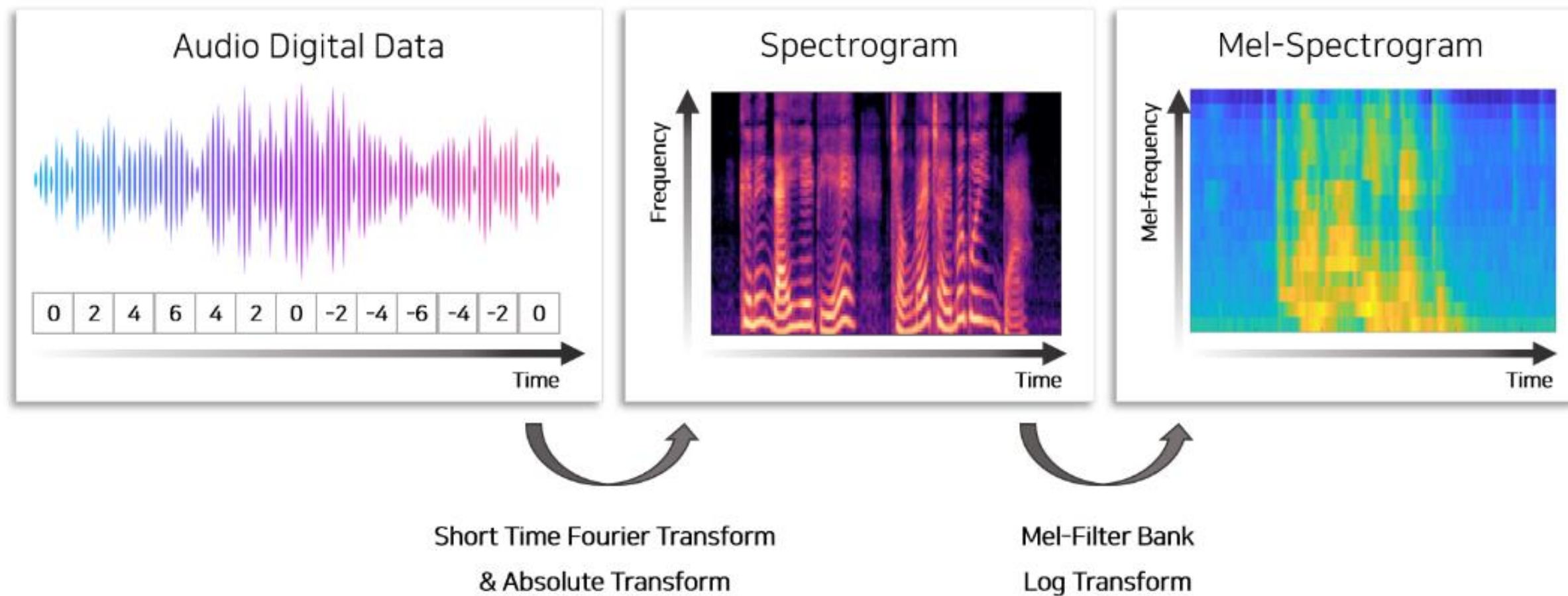
Model

<https://jounghEEKIM.github.io/2020/10/08/paper-review/>

*data processing

: 모델 학습을 위해 [input, label] 형태의 데이터가 필요

>> input : text (character) / label : mel-sepectrogram

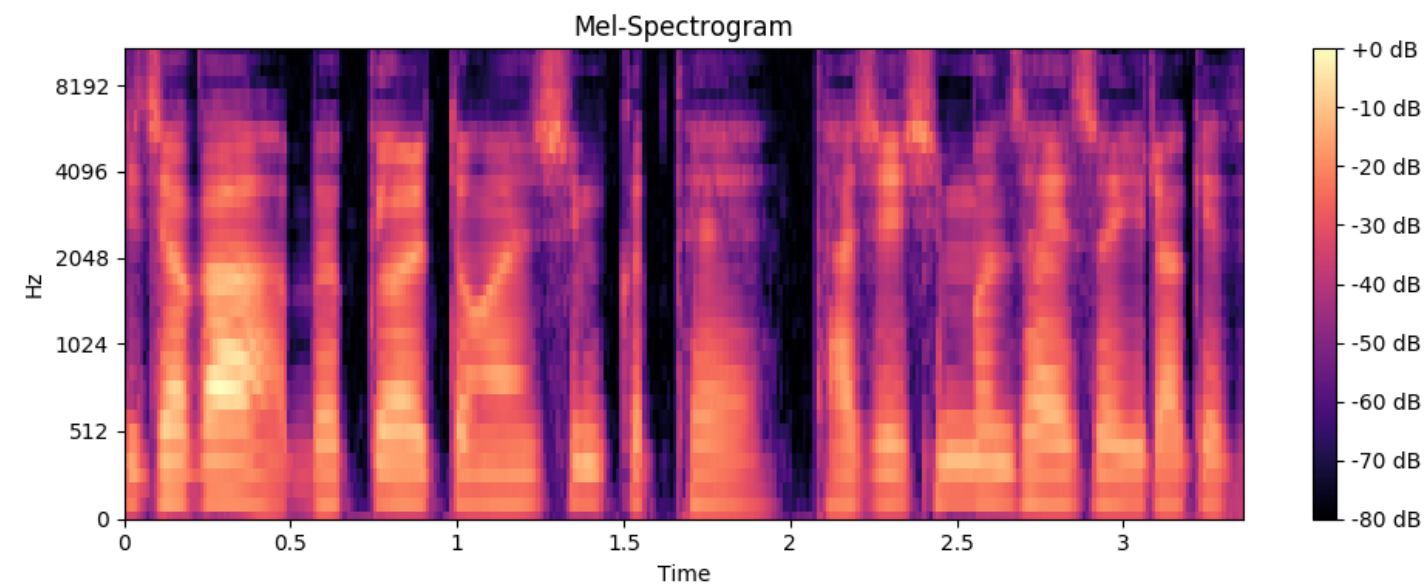


Model

* TTS (Text-to-Speech)

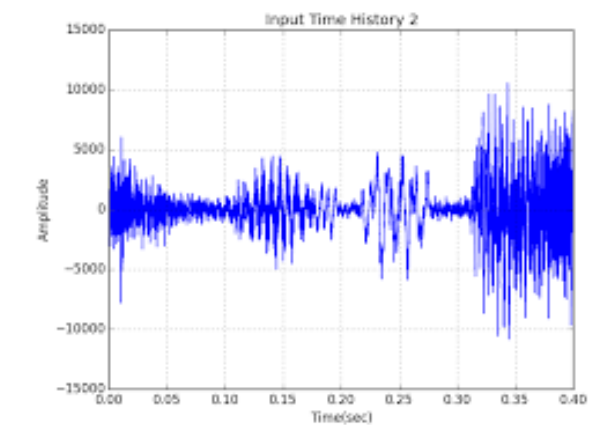
TEXT

text2mel



Mel-Spectrogram

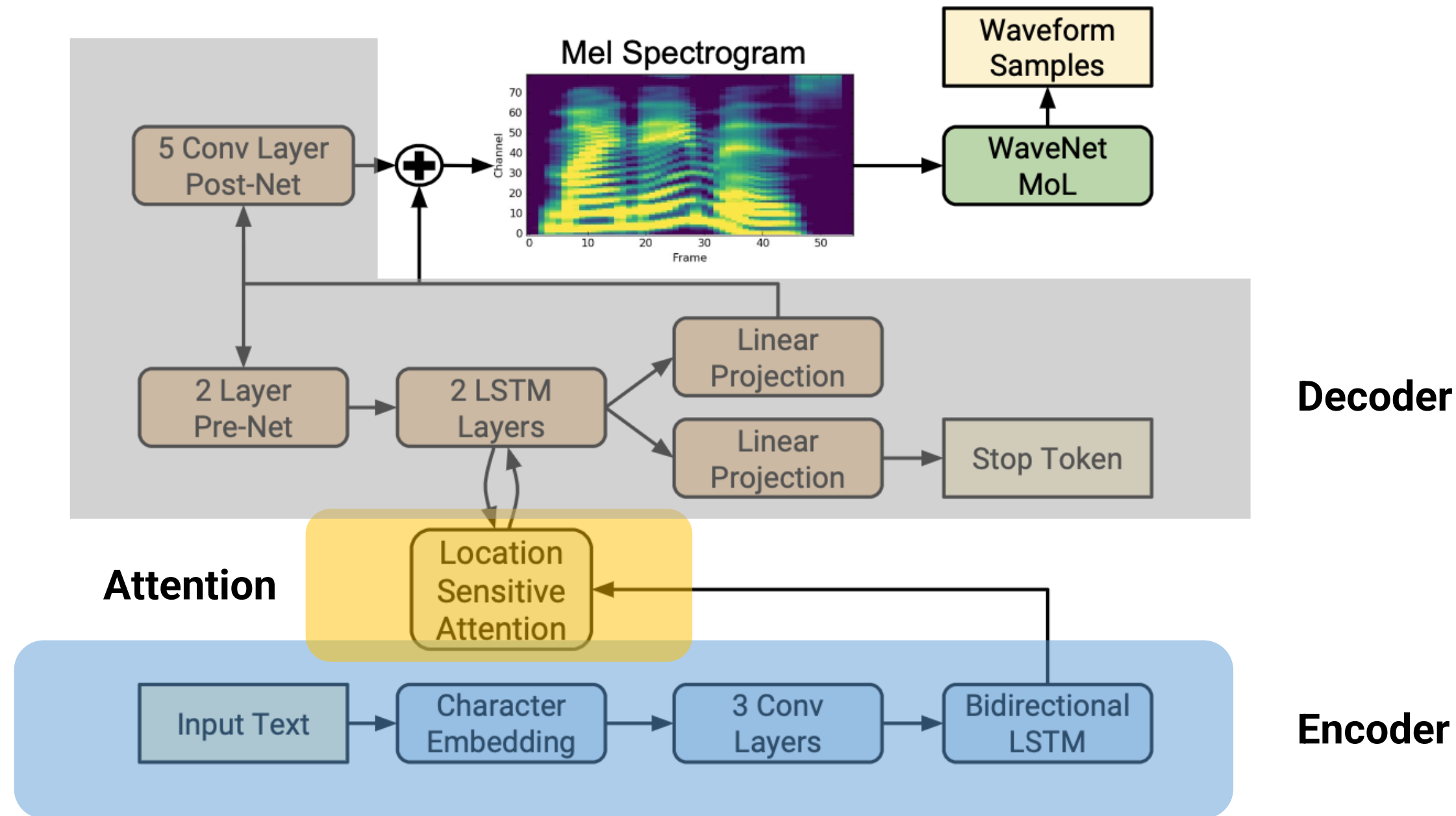
mel2wav



SPEECH

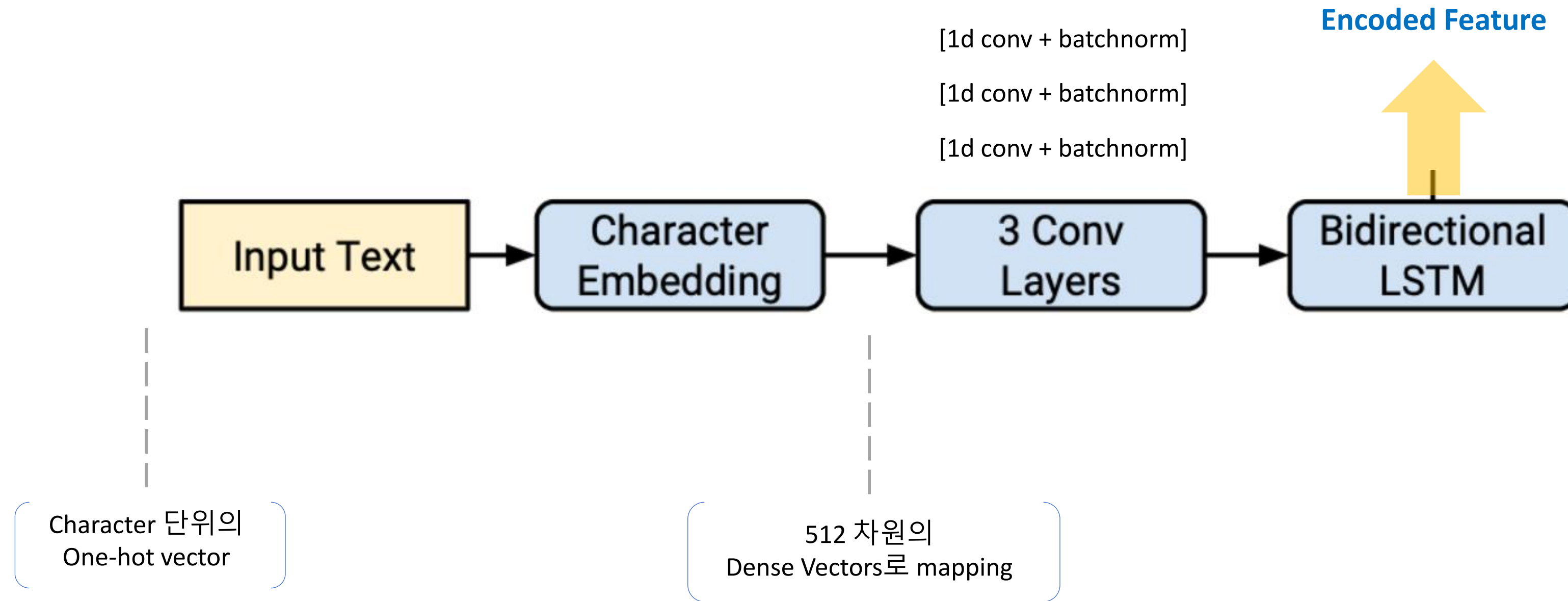
Model

* **text2mel** : Encoder & Decoder 구조



Model

* text2mel - Encoder

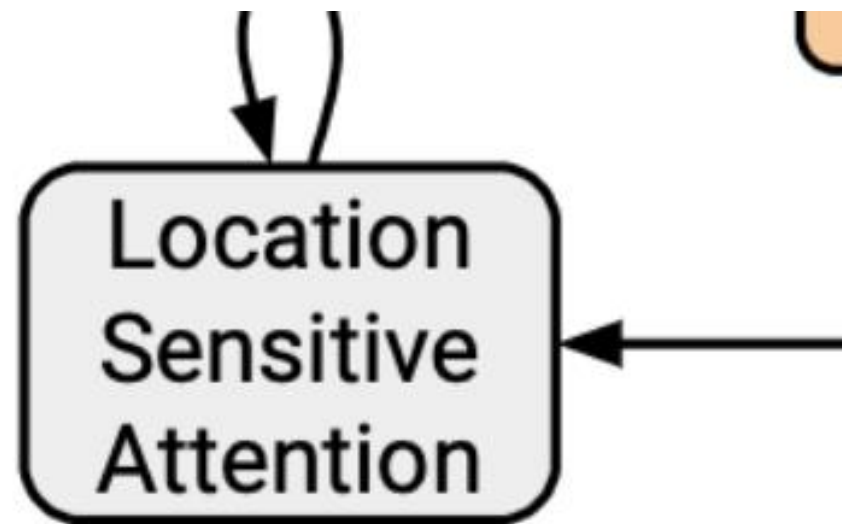


Model

* text2mel - Attention

: 매 시점 decoder에서 사용할 정보를 encoder에서 추출하여 가져오는 역할

: encoder에서 생성된 feature와 이전 시점의 mel-spectrogram을 이용하여 **encoder로부터 가져올 정보를 alignment**



Seq2seq 모델에서 흔히 사용되는 Bahdanau Attention에
Attention alignment 정보를 추가한 형태

$$e_i = W_a \tanh(W_b s_i + W_c h)$$
$$e_{i+1} = W_a \tanh(W_b s_i + W_c h + W_d e_i)$$

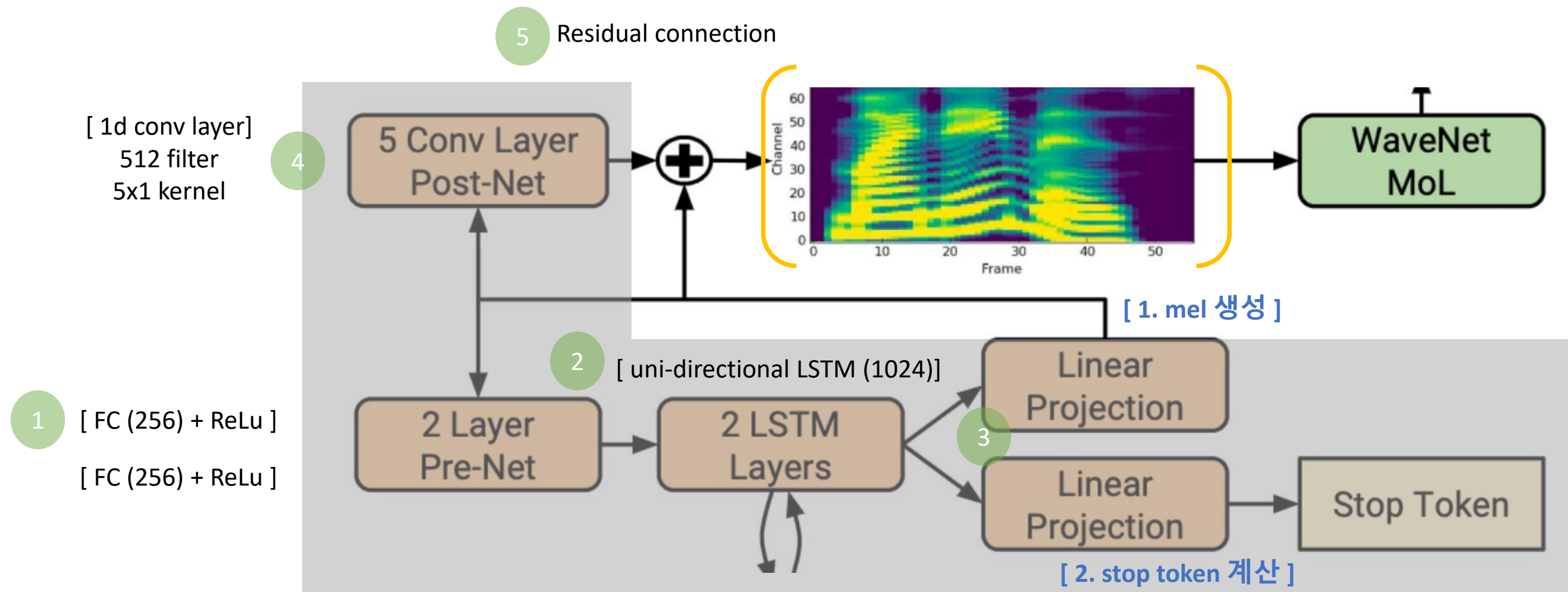
이전 step의 score를 추가해 자기회귀적(autoregressive) 으로
시간적 연관성(location sensitivity)을 고려하고자 고안

Attention 결과를 인접한 decoder step에서 consistent 하게 나타내기 위함

Model

* text2mel - Decoder

: alignment feature + 이전 시점 mel-spectrogram -> 다음 시점 mel-spectrogram 생성 / (loss function) = MSE



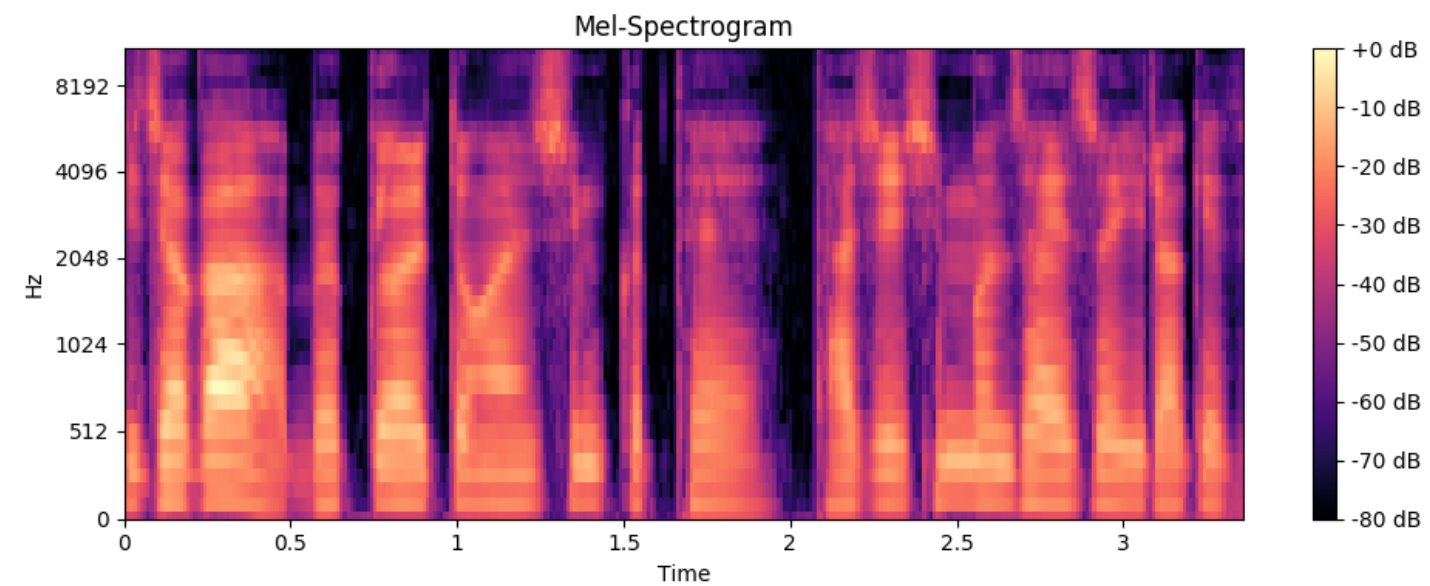
[FC layer] -> sigmoid

Model

* TTS (Text-to-Speech)

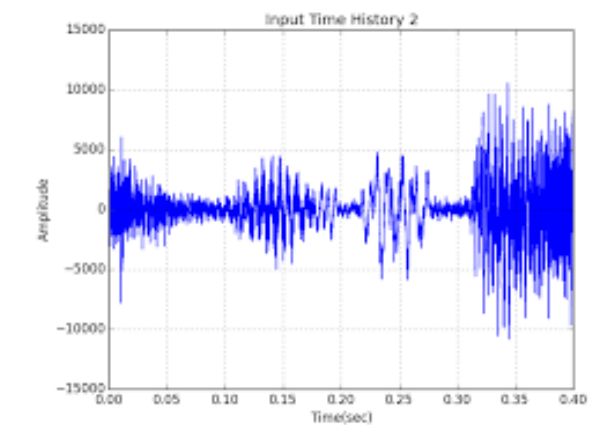
TEXT

text2mel



Mel-Spectrogram

mel2wav



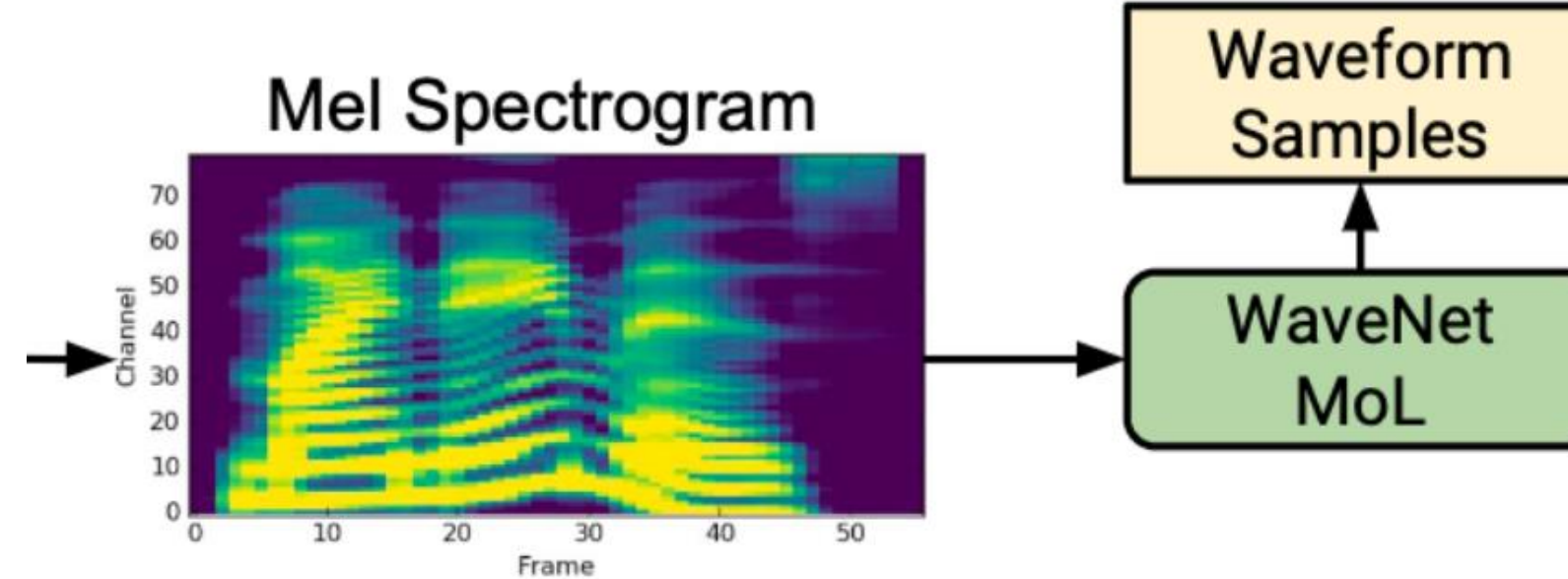
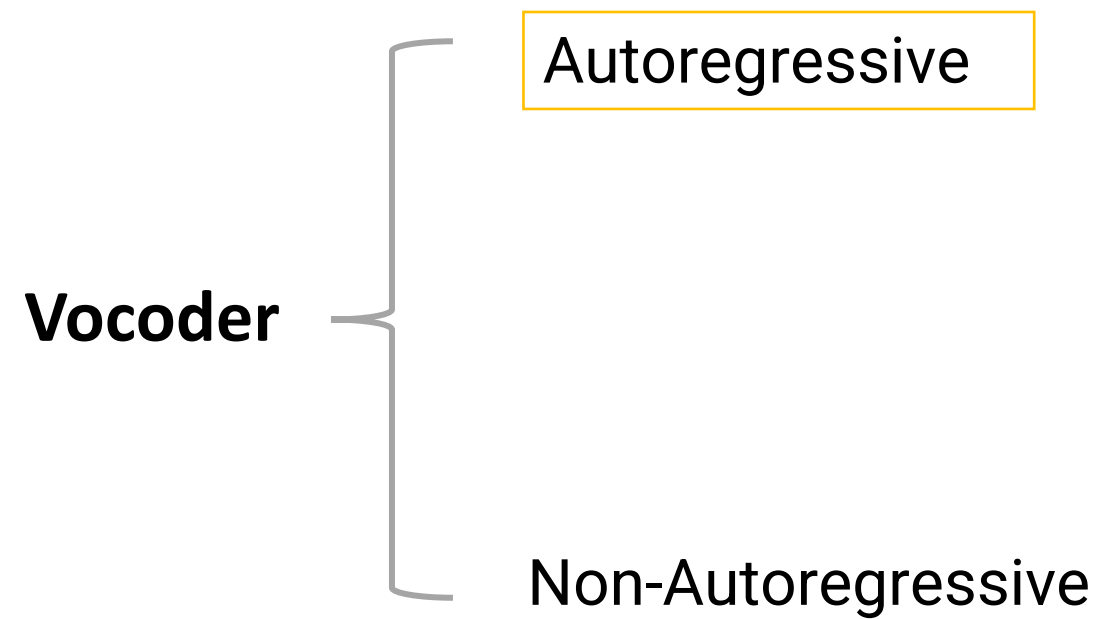
SPEECH

Model

* mel2wav – Vocoder (WaveNet)

: mel-spectrogram에서 wave form 생성 / 기존 wavenet의 구조 약간 변경 (MoL 차용)

: Encoder-Decoder 기반의 모델



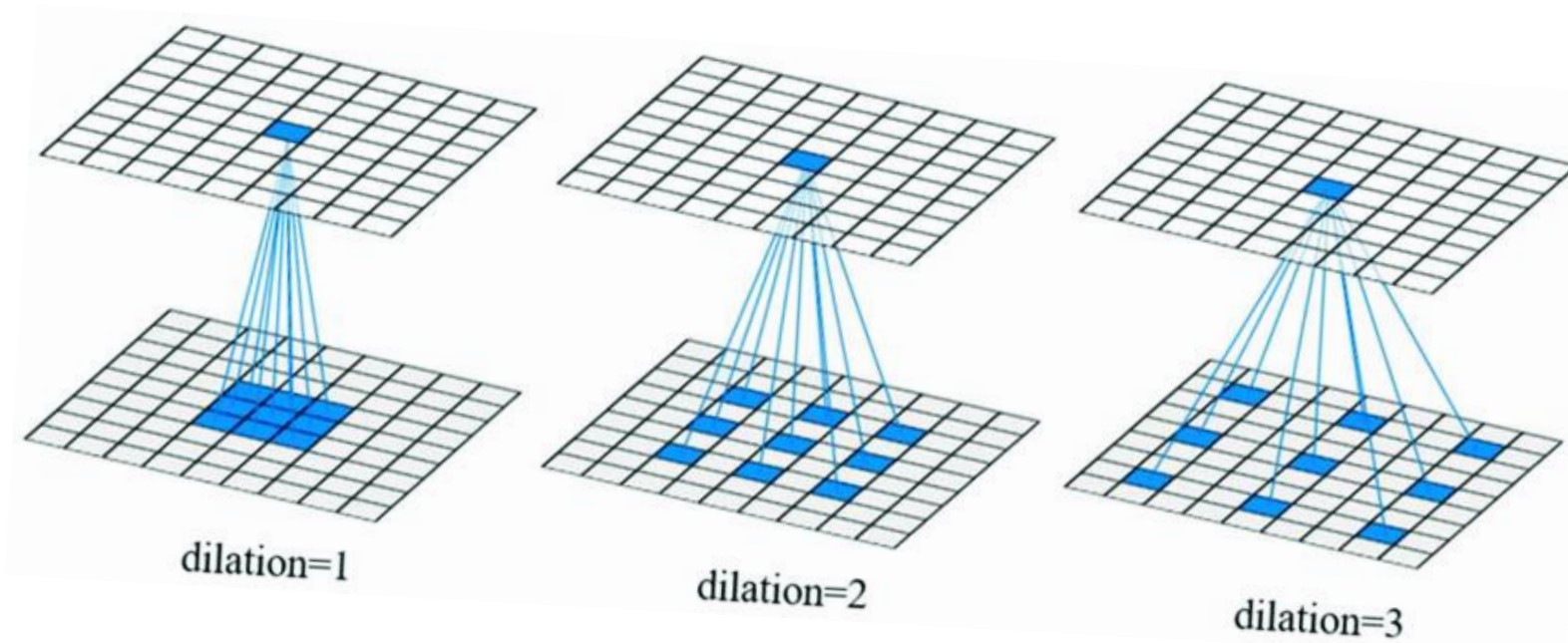
Model

* mel2wav – **Vocoder (WaveNet)**

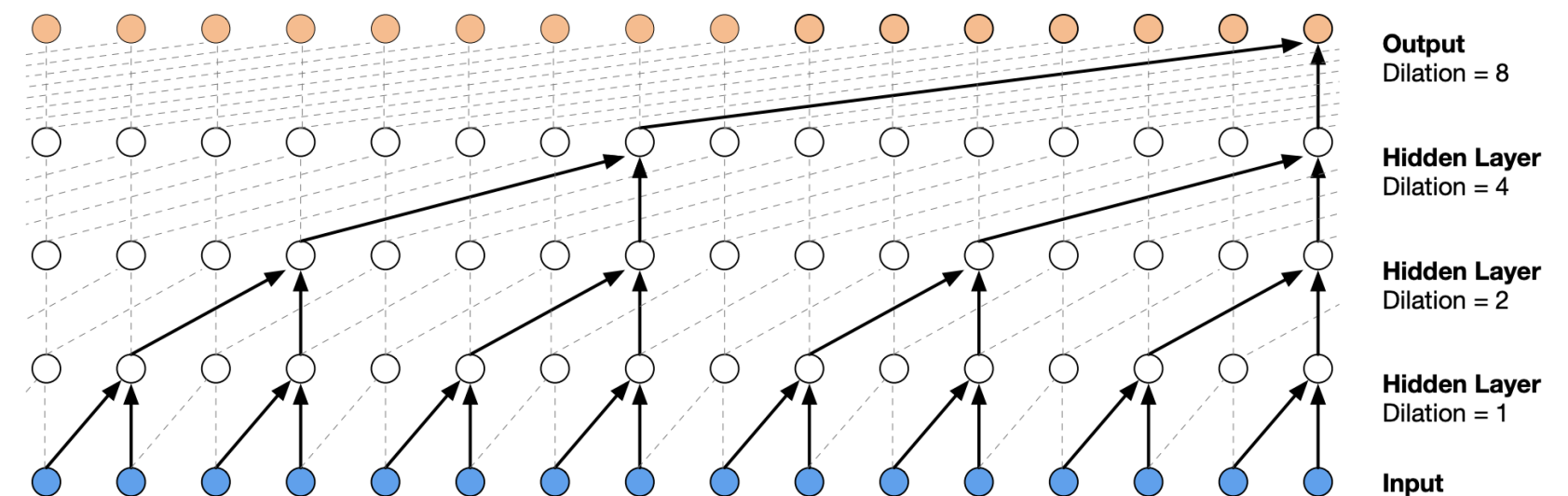
: Encoder-Decoder 기반의 모델

: Mel Spectrogram을 입력으로 받아 이를 **Dilated Causal Convolution**을 활용해 Encoding한 후 FCN과 Softmax로 Decoding하는 모델

: Randomized된 시작 wav 지점에서 n 개의 step 동안 전체 wav 파형을 예측



Dilated Causal Convolution



10 세트의 Dilated Causal Convolution

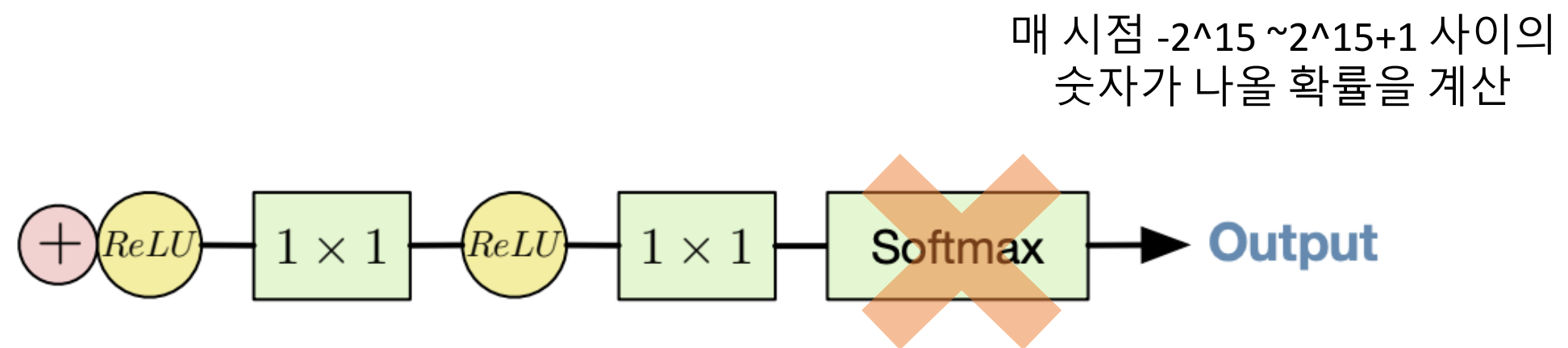
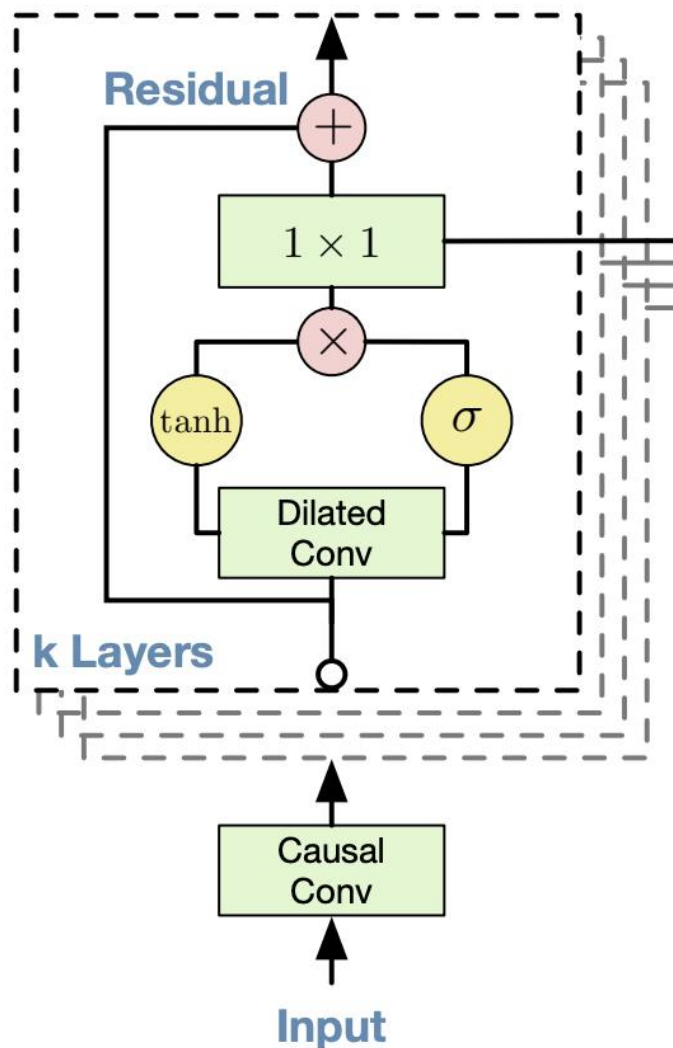
Model

* mel2wav – **Vocoder (WaveNet)**

: Encoder-Decoder 기반의 모델

: Mel Spectrogram을 입력으로 받아 이를 **Dilated Causal Convolution**을 활용해 Encoding한 후 FCN과 Softmax로 Decoding하는 모델

: Randomized된 시작 wav 지점에서 n 개의 step 동안 전체 wav 파형을 예측



Model

* mel2wav – **Vocoder (WaveNet)**

: **softmax**는 output node 간의 연관관계를 고려하지 못함 >> **MoL (Mixture of Logistic districution) 사용으로 변경**

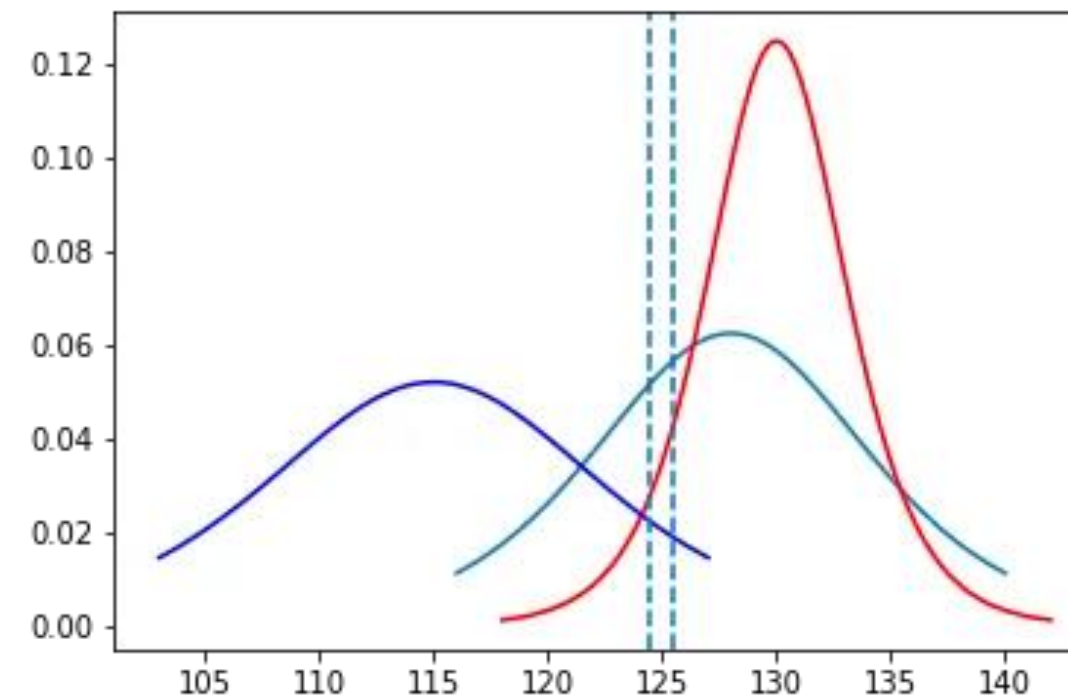
: Text2Mel의 최종 출력 영역은 총 128 차원짜리 Mel Frequency bin

-> 이 주파수 대역은 위치적 연관성이 곧 의미적 연관성을 보장 / 연속형 변수이기 때문 (회귀에 가까운 것)

: MOL은 Output이 여러 개의 Logistic Distribution의 가중합으로 이루어져 있다는 가정을 바탕으로 Output을 모델링

/ (loss function) = **Negative log-likelihood**

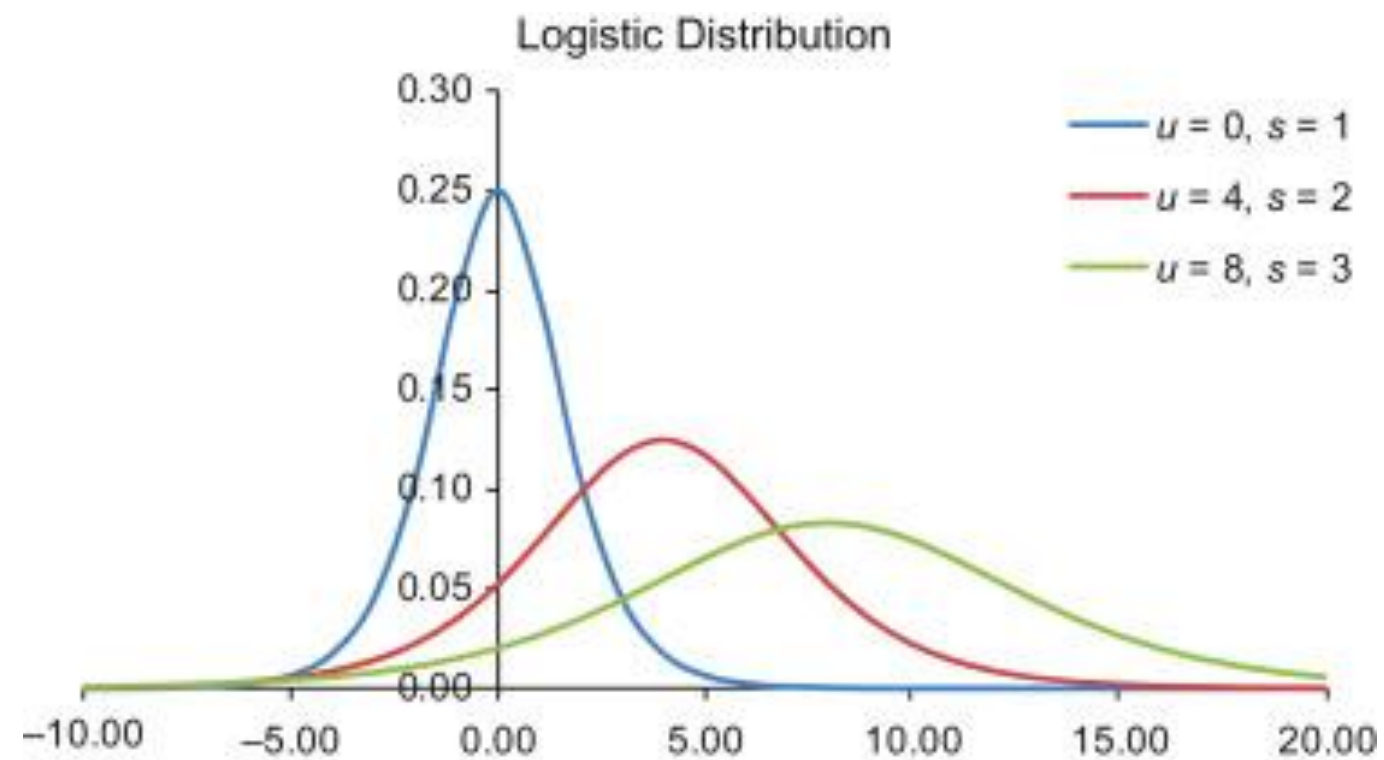
$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}$$
$$MOL(10) = a_1 f(x; \mu_1, s_1) + a_2 f(x; \mu_2, s_2) + \dots + a_{10} f(x; \mu_{10}, s_{10})$$
$$Output = \underset{x}{\operatorname{argmax}} MOL(10)$$



Model

* mel2wav – **Vocoder (WaveNet)**

+) Logistic distribution

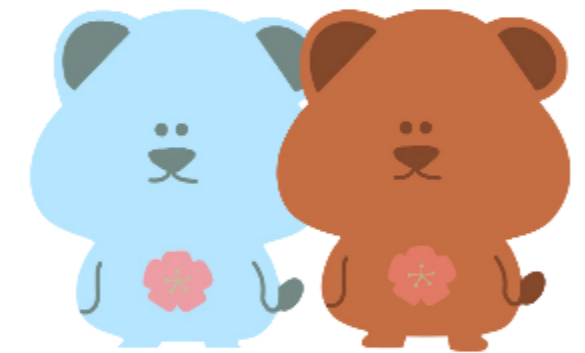


$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1 + e^{-(x-\mu)/s})^2}$$

$$F(x; \mu, s) = \frac{1}{1 + e^{-s(x-\mu)}}$$

Mu = 0, s =1로 두면 logistic function == sigmoid

Model Training



Model Training

* Training Setup

teaching-forcing 사용

* text2mel : Train 시에는 이전 시점에서 생성된 mel-spectrogram 이 아니라

Ground-truth mel-spectrogram을 사용하여 학습 효율 증가

* mel2wav : input으로 ground-truth waveform 이용

Evaluation



Evaluation

* metric? -> **MOS**

Mean Opinion Score(MOS)는 사회과학과 공학 전반에서 활용되는 모델 평가 기준 중 하나
모델이나 시스템의 성능을 실험하고 체험자가 1점부터 5점까지 5개의 만족도 중 하나 선택
→ 여러 체험자의 값을 평균 / 5점에 가까울수록 모델의 성능이 뛰어나다.

자연스러운 음성합성은 다양함

해당 텍스트에 대한 자연스러운 음성은 무궁무진하게 많고
따라서 이러한 **주관성을 반영**할 수 있는 평가 지표로 MOS가 활용되는 것

Evaluation

24.6 시간 동안 한사람의 음성을 담은 US English dataset 이용
음성을 들려주고 1점~5점 중 0.5 간격으로의 점수를 주도록 함 (MOS)

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	4.526 ± 0.066

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

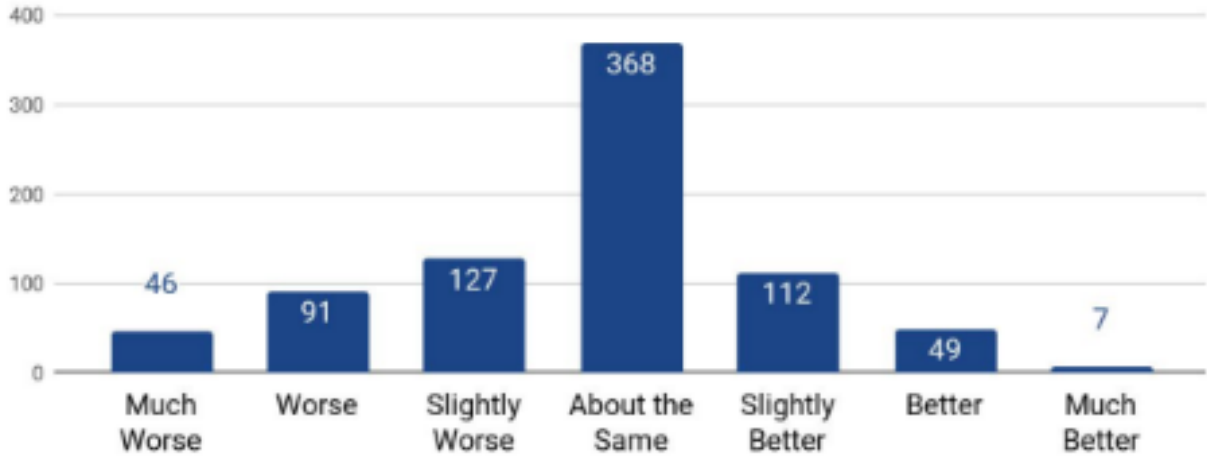


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

Evaluation

(1) Predicted Features versus Ground Truth

Training	Synthesis	
	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

Table 2. Comparison of evaluated MOS for our system when WaveNet trained on predicted/ground truth mel spectrograms are made to synthesize from predicted/ground truth mel spectrograms.

(2) Linear Spectrograms

System	MOS
Tacotron 2 (Linear + G-L)	3.944 ± 0.091
Tacotron 2 (Linear + WaveNet)	4.510 ± 0.054
Tacotron 2 (Mel + WaveNet)	4.526 ± 0.066

Table 3. Comparison of evaluated MOS for Griffin-Lim vs. WaveNet as a vocoder, and using 1,025-dimensional linear spectrograms vs. 80-dimensional mel spectrograms as conditioning inputs to WaveNet.

(3) Simplifying WaveNet

Total layers	Num cycles	Dilation cycle size	Receptive field (samples / ms)	MOS
30	3	10	6,139 / 255.8	4.526 ± 0.066
24	4	6	505 / 21.0	4.547 ± 0.056
12	2	6	253 / 10.5	4.481 ± 0.059
30	30	1	61 / 2.5	3.930 ± 0.076

Table 4. WaveNet with various layer and receptive field sizes.

Conclusion



Conclusion

sequence-to-sequence recurrent network with attention
to predicts mel spectrograms with a modified **WaveNet vocoder**

WaveNet을 이용해서 음성 합성 퀄리티 상승
복잡한 feature engineering 불필요

SOTA 달성!

THANK YOU

