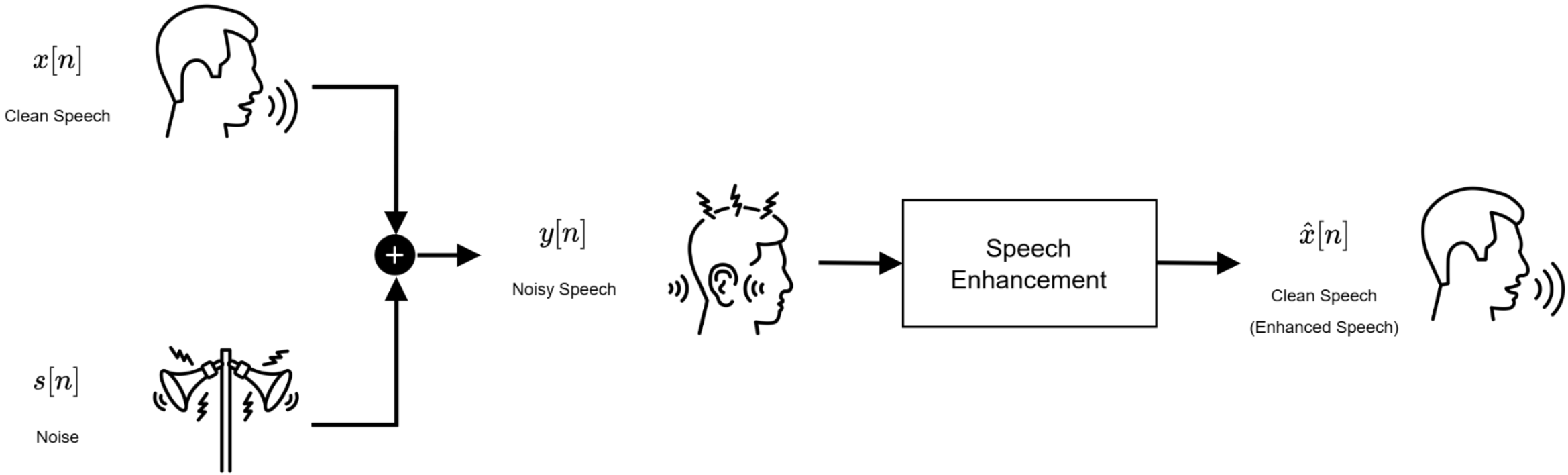
[졸업작품전 – 작품]

Lightweight Real-time Speech Enhancement with Temporal Convolutional Neural Networks

- 김 다빈 : 융합전자공학부 2021015478
- 김 태희 : 융합전자공학부 2019015014

Task 소개: Speech Enhancement

- 음향 노이즈의 영향을 받은 음성 신호를 깨끗한 음성 신호로 복구하는 과제



목표 소개: Lightweight Real-time

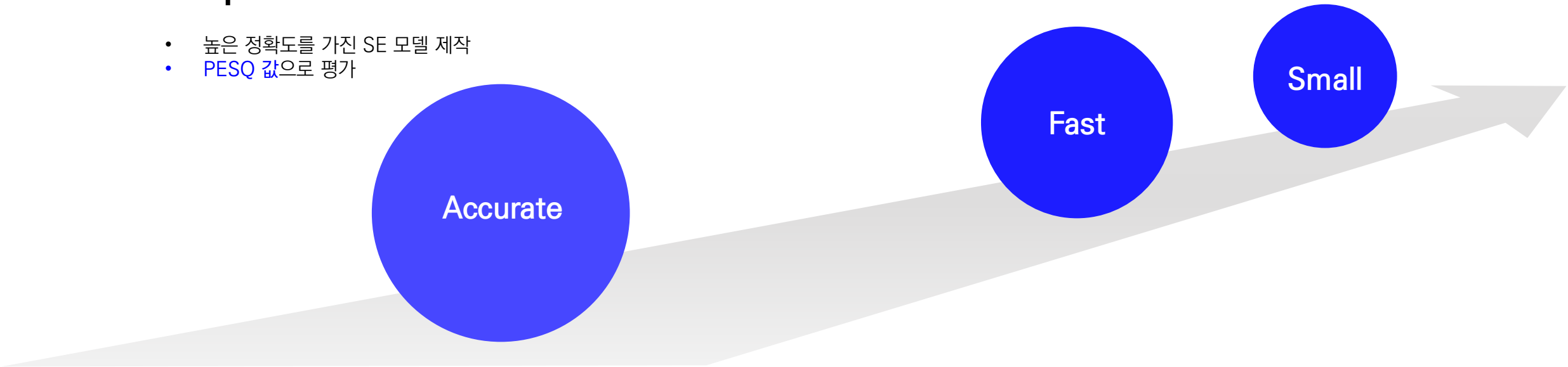
- 원격회의, 보청기 등 다양한 분야의 전처리기로써 사용
  - ➢ 명확하게 잡음이 제거되면서, 모델이 가볍고 빠르게 경량화 되는 것이 중요

Step1

- 높은 정확도를 가진 SE 모델 제작
- PESQ 값으로 평가

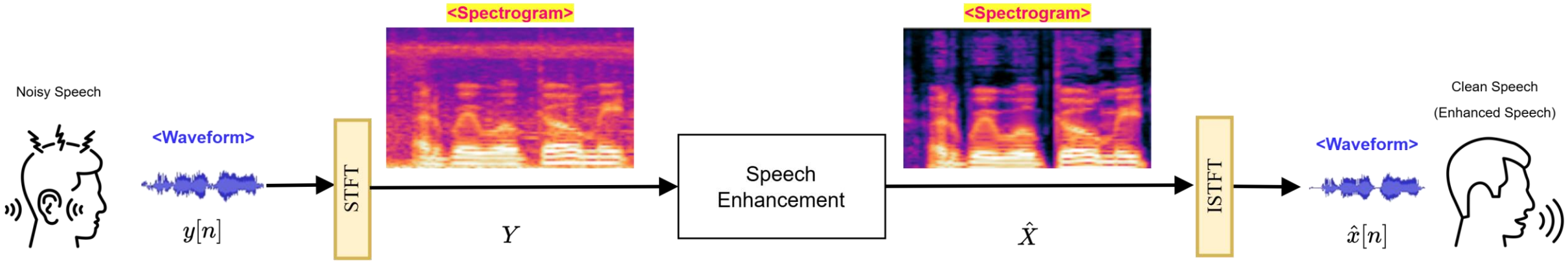
Step2

- 빠른 것은 RTF를 통해 평가
- 작은 것은 파라미터 수로 평가



PESQ (Perceptual Evaluation of Speech Quality): 음성 품질을 인간의 청각 시스템 관점에서 평가하기 위해 개발된 알고리즘으로 실제 청취자가 느낄 품질에 가까운 점수를 제공
RTF (Real Time Factor): 음성 처리 시스템에서 처리 속도를 평가하는 지표, 처리시간/실제 음성 길이로 정의 됨.

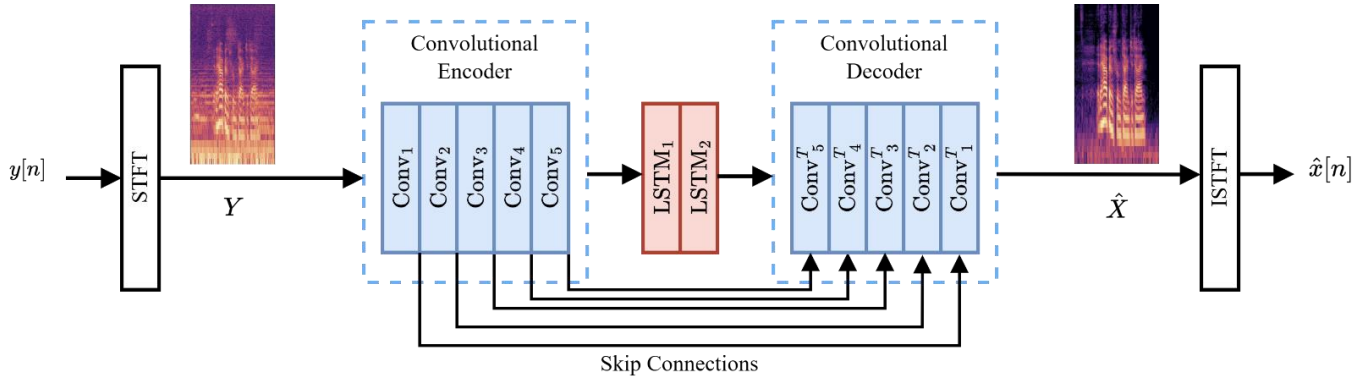
Speech 데이터의 특성: Sequential, Spectrogram

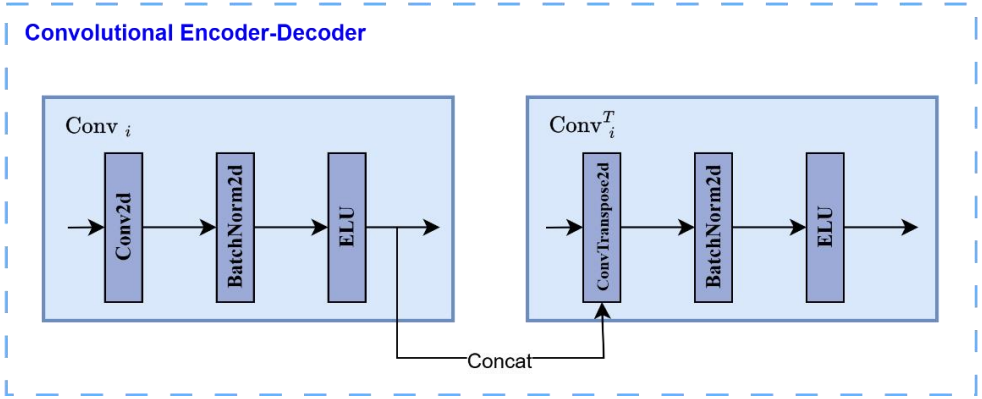


- Spectrogram은 waveform을 작은 시간 구간으로 나누어, 각 구간에 대해 Fourier Transform을 한 것을 시각화 한 자료

- 이미지 데이터의 특성과 시간에 따라 변화하는 Sequential 데이터의 특징을 동시에 가짐
  - ✓ CNN으로 인코더와 디코더를 구성하여 Spectrogram의 공간적 정보를 고려
  - ✓ LSTM을 인코더와 디코더 사이에 넣어 Sequential 정보를 고려한 모델을 Baseline으로 선정

Baseline 구조 소개: CED + LSTM





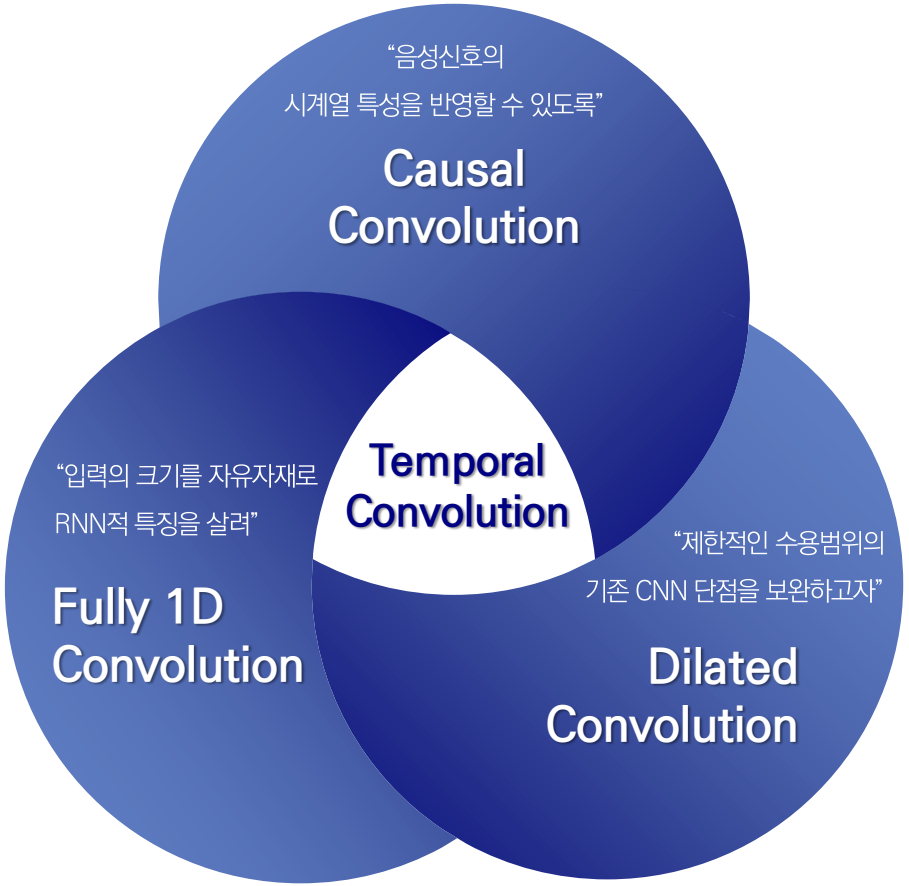
	Parameters	Inference Time(sec)
CNN Encoder	262,704 (1.5%)	0.075
LSTM	16,793,600 (95.5%)	1.634
CNN Decoder	523,155 (3%)	0.187
Total	17,579,459	1.903

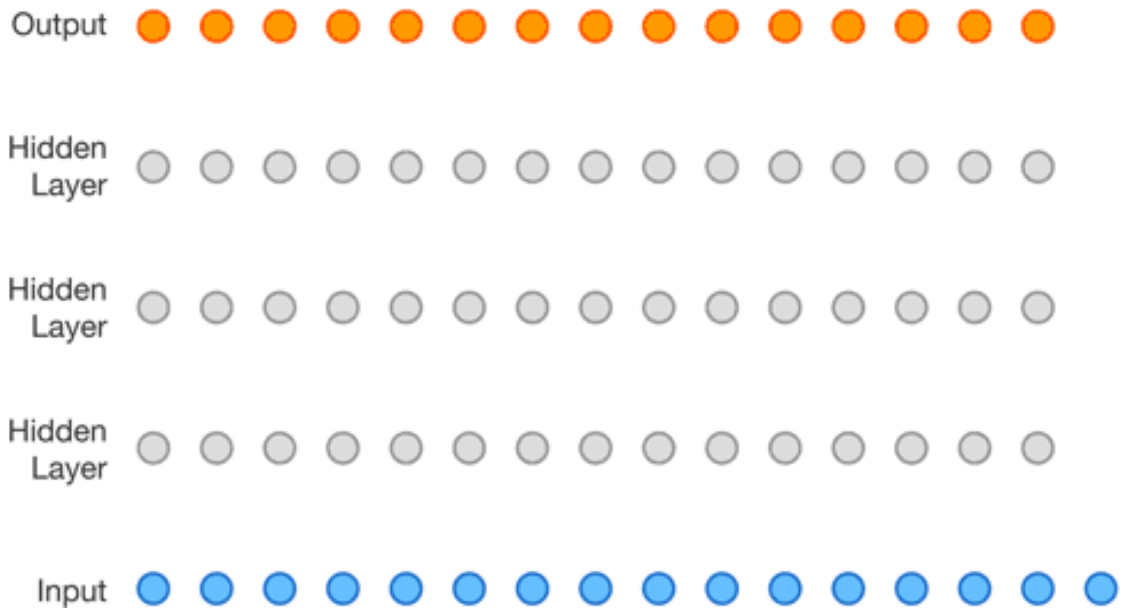
➢ 전체 파라미터의 약 95%, 전체 Inference time의 약 86%를 차지하는 LSTM

- 이미지 데이터의 특성과 시간에 따라 변화하는 Sequential 데이터의 특징을 동시에 가짐
  - ✓ CNN으로 인코더와 디코더를 구성하여 Spectrogram의 공간적 정보를 고려
  - ✓ LSTM을 인코더와 디코더 사이에 넣어 Sequential 정보를 고려한 모델을 Baseline으로 선정

Temporal Convolutional Network (TCN)

- LSTM의 특징을 살리면서, 경량화를 가능하게 할 **Temporal Convolutional Network !!**



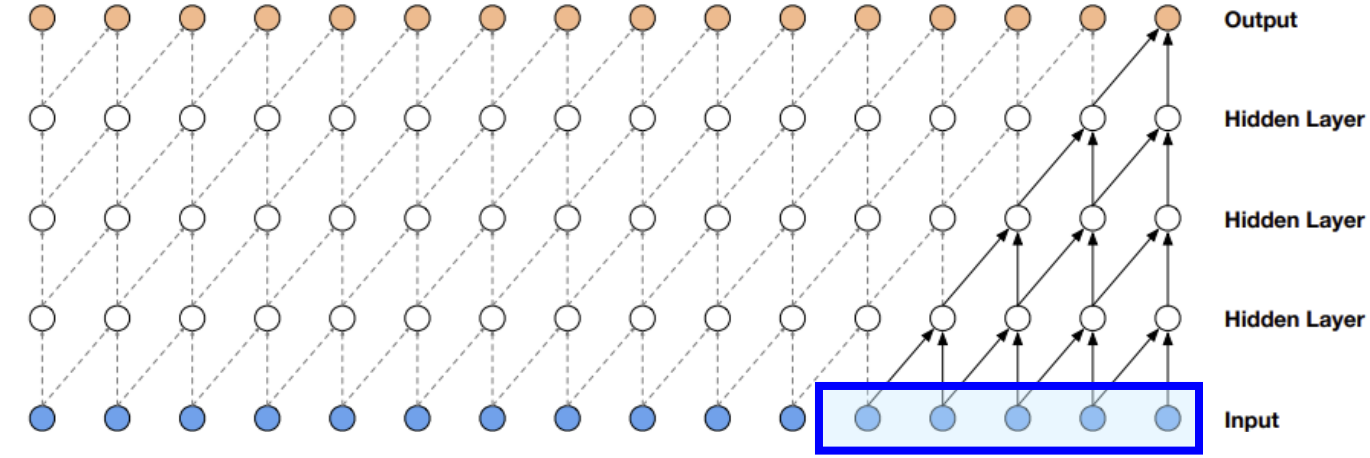


- ✓ Causal: 현재의 출력이 미래의 입력에 영향을 받지 않음
- ✓ Fully 1D: 입력의 크기에 무관하게 출력 가능
- ✓ Dilated: Receptive Field가 제한적이었던 기존 CNN의 문제를 해결

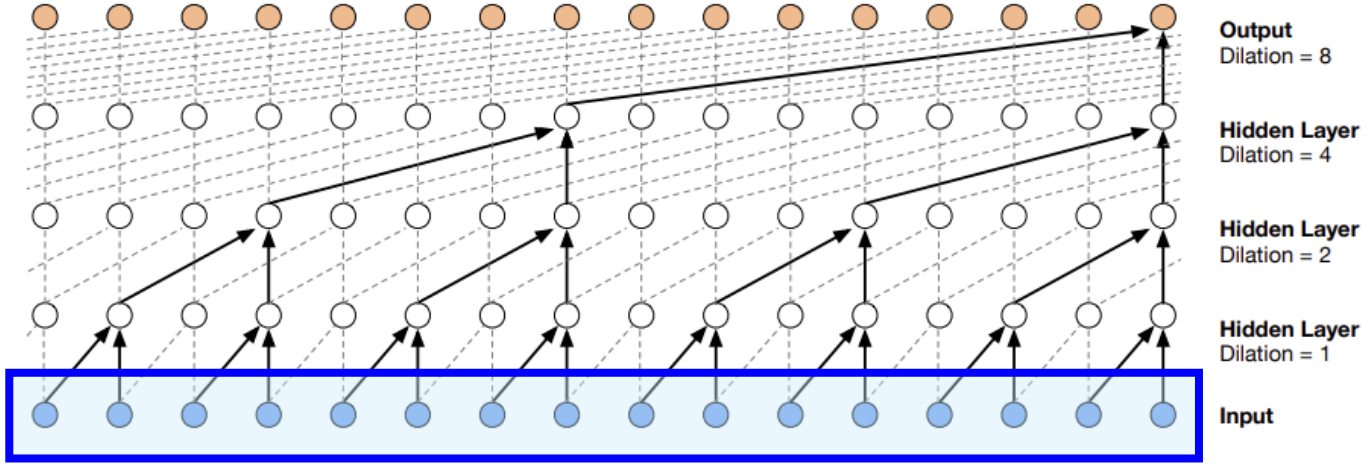
Dilated Convolution

- Dilation rate을 조정하여 Receptive Field를 늘렸음

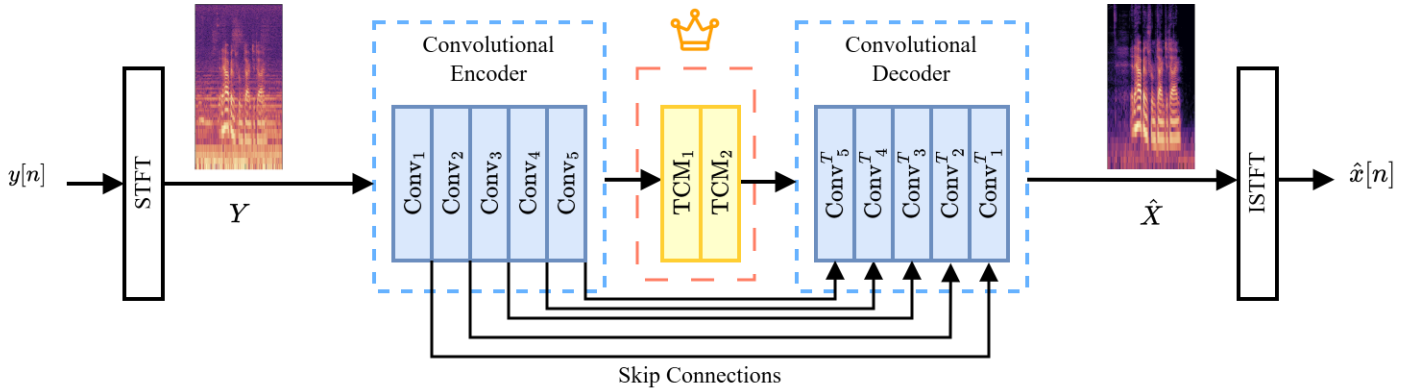
➢ Causal Convolution

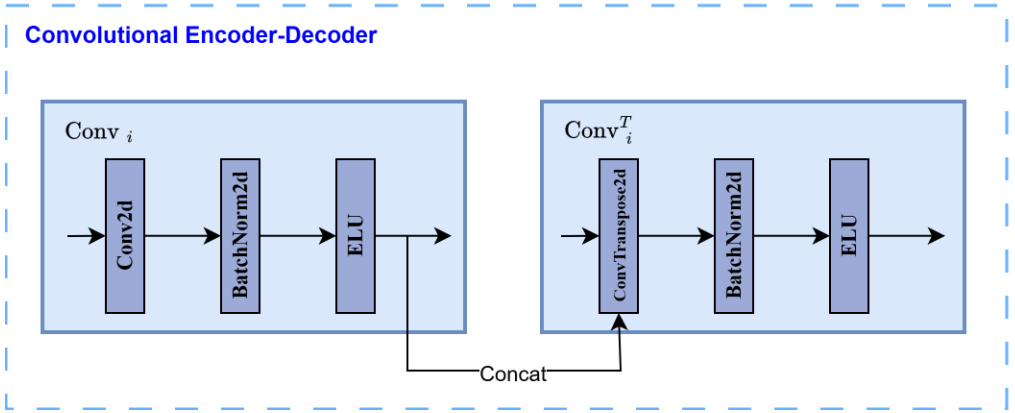


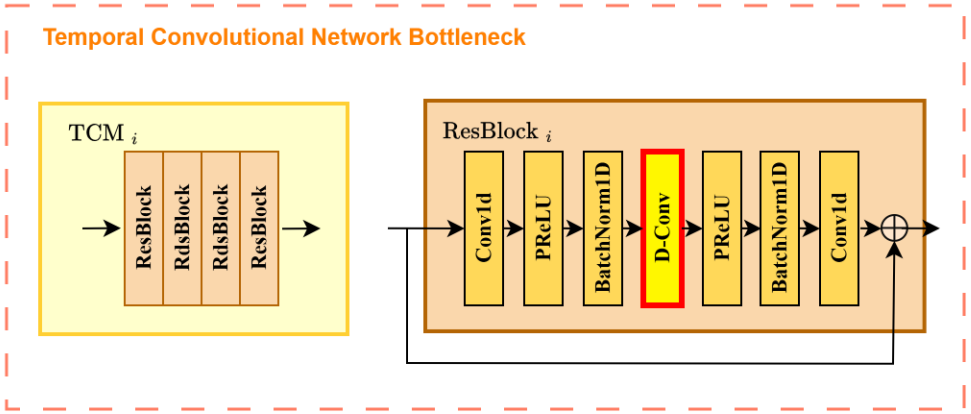
➢ Causal Dilated Convolution



Temporal Convolutional Neural Network







- Encoder, Decoder 사이에 2개의 TCM을 sequential하게 연결
- Dilated Convolution Layer를 포함하고 있는 Residual Block을 직렬로 이어 붙인 구조
- Dilation의 값을 2의 제곱수로 키워가며 receptive field를 넓힌 구조
- Residual Block을 이용해 Residual Learning이 가능하도록 함

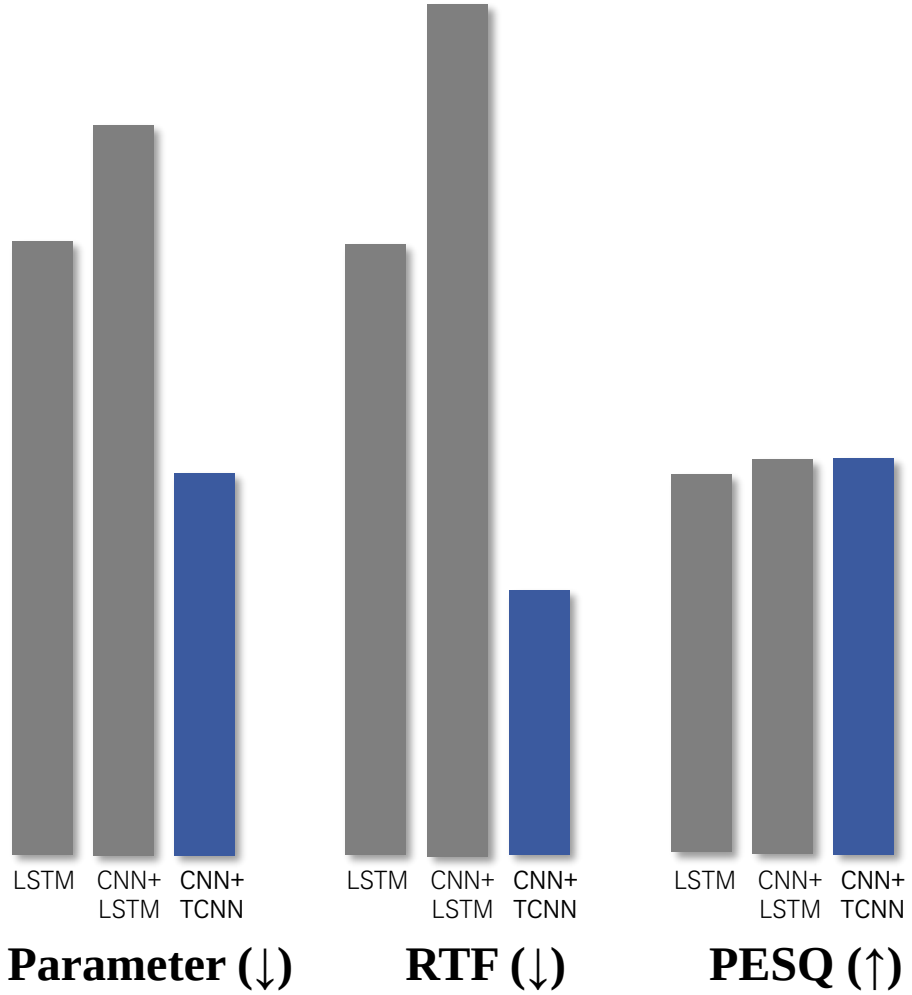
LSTM vs TCM의 Parameter, Inference time 상세 비교

	Parameters	Inference Time(sec)
CNN Encoder	262,704	0.075
LSTM	16,793,600	1.634
CNN Decoder	523,155	0.187
Total	17,579,459	1.903

	Parameters	Inference Time(sec)
CNN Encoder	262,704	0.079
TCM	8,421,392	0.289
CNN Decoder	523,155	0.262
Total	9,207,251	0.634

- Baseline의 LSTM 대비 TCM Parameter 수 약 50% 감소, Inference Time 약 82% 감소

Result: 전체 모델 구조 분석



	LSTM	CNN+LSTM (Baseline)	CNN+TCM (Proposed)
Parameter (↓) 얼마나 작은가	14.76×10^{6}	17.58×10^{6}	$\mathbf{9.21\times10^{6}}$
RTF (↓) 얼마나 빠르게 출력되나	20.2×10^{-2}	28.2×10^{-2}	$\mathbf{8.79\times10^{-2}}$
PESQ (↑) 얼마나 정확한가	2.5375	2.6087	**2.6209**

- Baseline 대비 Parameter 수 약 48% 감소, RTF값 약 69% 감소, PESQ값 약간 상승
  - PESQ: 테스트셋 (393개)의 평균 PESQ값 측정
  - RTF: 테스트셋 (393개)의 평균 RTF (Inference Time / Total Audio Length)
  - Parameter: Total Trainable Parameter의 개수

Demo

	Sample 1	Sample 2	Sample 3	Sample 4
Noisy 				
Enhanced 				
Baseline Model Inference Time	2.00s	2.00s	2.00s	2.00s
TCNN Inference Time	0.47s	0.34s	0.21s	0.46s

Q&A
•
Appendix

실험 상세정보

- Dataset: VoiceBank + DEMAND
- Loss: Mean Square Error
- Sampling rate: 16 kHz
    - Frame Size: 20 ms, Frame shift: 10 ms

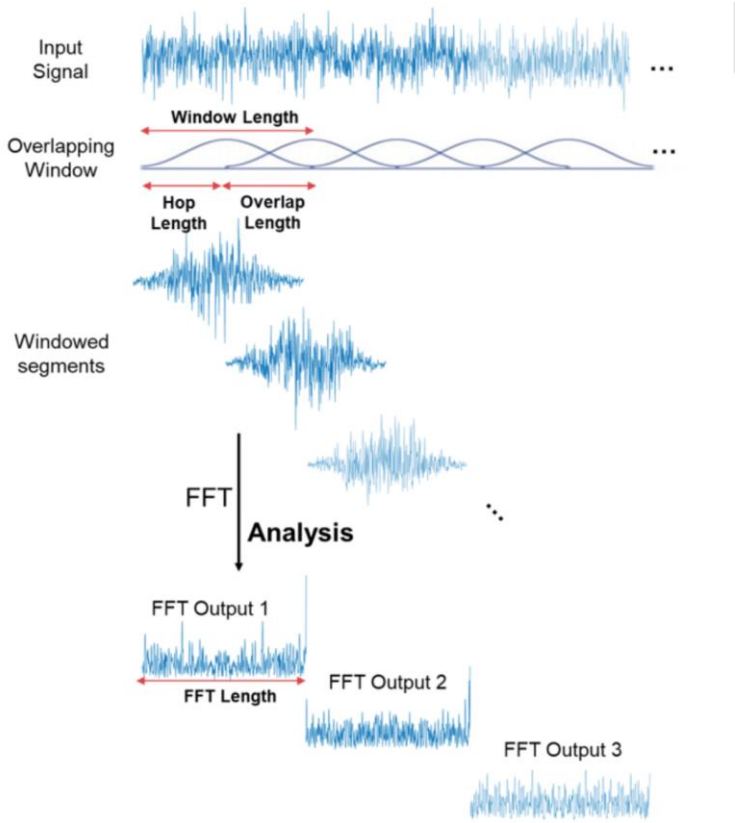
- Optimizer: Adam
- Learning rate: 0.002
- Batch size: 16
- Epoch: 100에서 Best 선택

Short Time Fourier Transform (STFT)

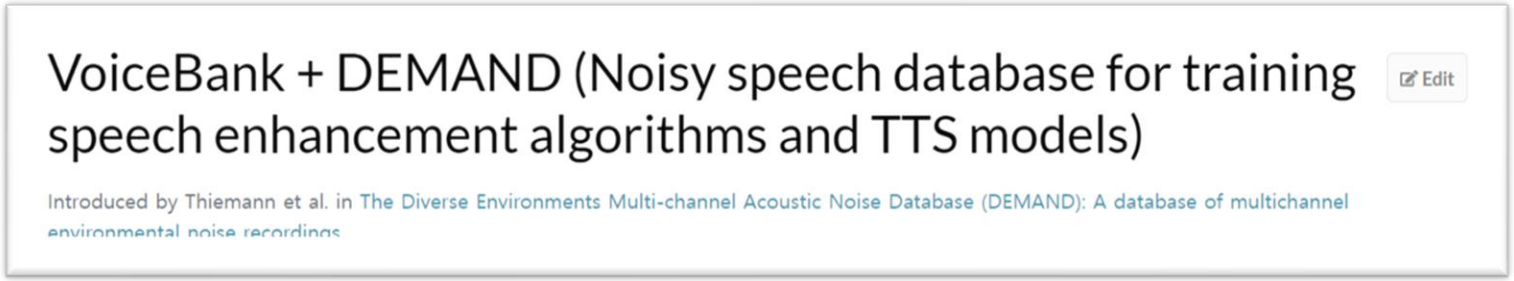
- Time domain에서 Frequency domain으로 매핑 시켜주는 역할을 하는 Fourier Transform
  ⇒ 시간 도메인의 정보를 완전히 잃어버린다는 단점이 있음

- Short Time Fourier Transform (STFT)
  - Time window를 움직이며 Fourier Transform
  - Frame 간 정보가 자연스럽게 이어지게 하기 위해 overlap을 시킴
  - 시간 frame마다 [주파수x세기] 데이터를 얻을 수 있음
  - 시간을 x축, 주파수를 y축, 세기를 magnitude로 시각화하면 Spectrogram을 얻을 수 있음

- 코드에서의 설정값: Sample 기준
  - Time window length : 320 (20 ms)
  - Hop length : 160 (10 ms)



Dataset: VoiceBank+DEMAND



- VoiceBank
  - VoiceBank는 Clean 음성 데이터셋
- VoiceBank+DEMAND
  - SNR(음성 데이터에 추가된 잡음의 양)은 0dB, 5dB, 10dB, 15dB로 구성
  - 약 12시간의 데이터
  - 학습 데이터는 28명의 발화자, 평가 데이터는 2명의 발화자로 구성
    - 평가 데이터 2명을 각각 Validation과 Test로 나누어서 사용하였음
- DEMAND
  - Diverse Environments Multichannel Acoustic Noise Database
  - DEMAND는 다양한 일상 환경의 배경 소음을 수집한 데이터베이스
    - 예: 사무실, 카페, 지하철 등

MSE Loss 상세

- 마스크 생성
  - 각 배치의 유효한 프레임 수 만큼의 마스크를 생성하여 해당 프레임까지만 손실에 포함, 패딩이나 잡음만 있는 구간은 0으로

$$M_i\,(t,d) = \begin{cases} 1, if\ t < n\ frames[i] \\ 0, otherwise \end{cases}$$

- 마스킹된 예측 및 레이블 ($\mathbb{R}^{B\times T\times D}$, *B: 배치 사이즈, T: 최대 시간 길이, D: 특징 차원*)

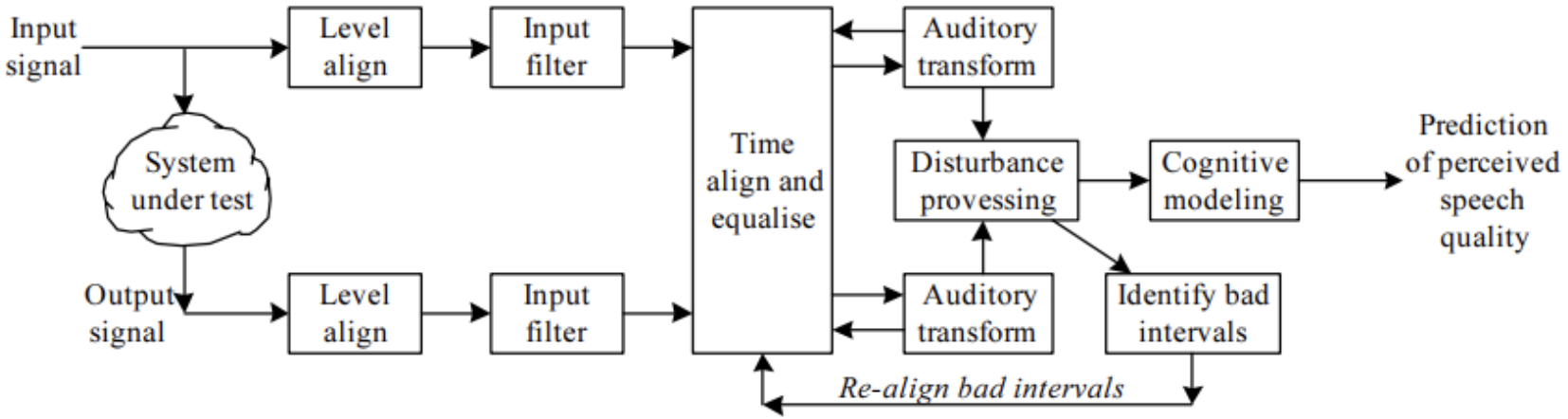
$$\hat{X}_{masked} = \hat{X} \odot M$$
$$X_{masked} = X \odot M$$

- 손실함수

$$Loss = \frac{\sum_{i=1}^{B}\sum_{t=1}^{T}\sum_{d=1}^{D}(\hat{X}_{masked,i,t,d} - X_{masked,i,t,d})^2}{\sum_{i=1}^{B}\sum_{t=1}^{T}\sum_{d=1}^{D} M_{i,t,d} + \epsilon}$$

PESQ

- Perceptual Evaluation of Speech Quality (PESQ)
  - 음성 품질의 주관적 평가방법과 비슷한 결과를 내면서 객관적으로 성능을 측정할 수 있도록 만들어진 평가 방법이며, 음성품질을 **인간의 청각 시스템 관점에서 측정**한 것
  - 입력 음성신호와 시스템을 통과한 출력 음성신호를 표준청각 레벨을 기준으로 정렬 → 수화기의 대역통과 특성을 고려하기 위한 필터링 과정 → 시간정렬, 비가청음 영역의 신호 제거 → 사람의 청각특성을 고려한 지각영역으로의 변환 → 인지과정 모델



Parameter 수 계산

- Conv & Transpose Conv 파라미터 수 계산 방식

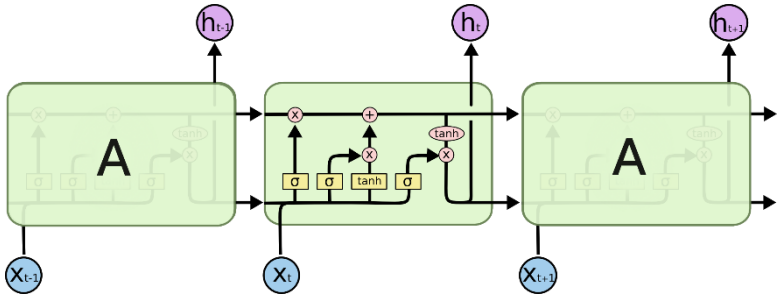
$$weight + bias =$$

$$[\ \#output\ channel\ ] * [\ \#input\ channel\ ] * [\ kernel\ size\ ] + [\ \#output\ channel\ ]$$

- LSTM 파라미터 수 계산 방식

$$weight + bias =$$

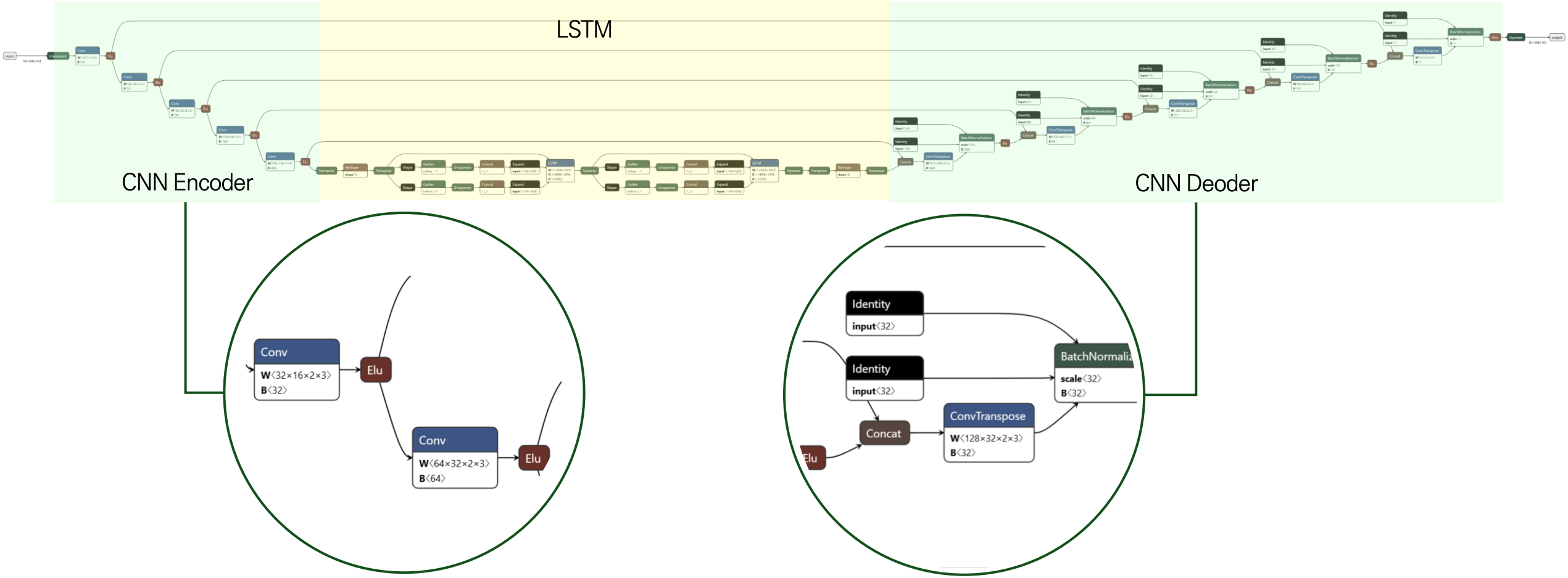
$$4\ *\ [\ hidden\ size\ ] * [\ input\ size\ ] + 4\ *\ [\ hidden\ size\ ] * [hidden\ size\ ] + 8\ *\ [\ hidden\ size\ ]$$



$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

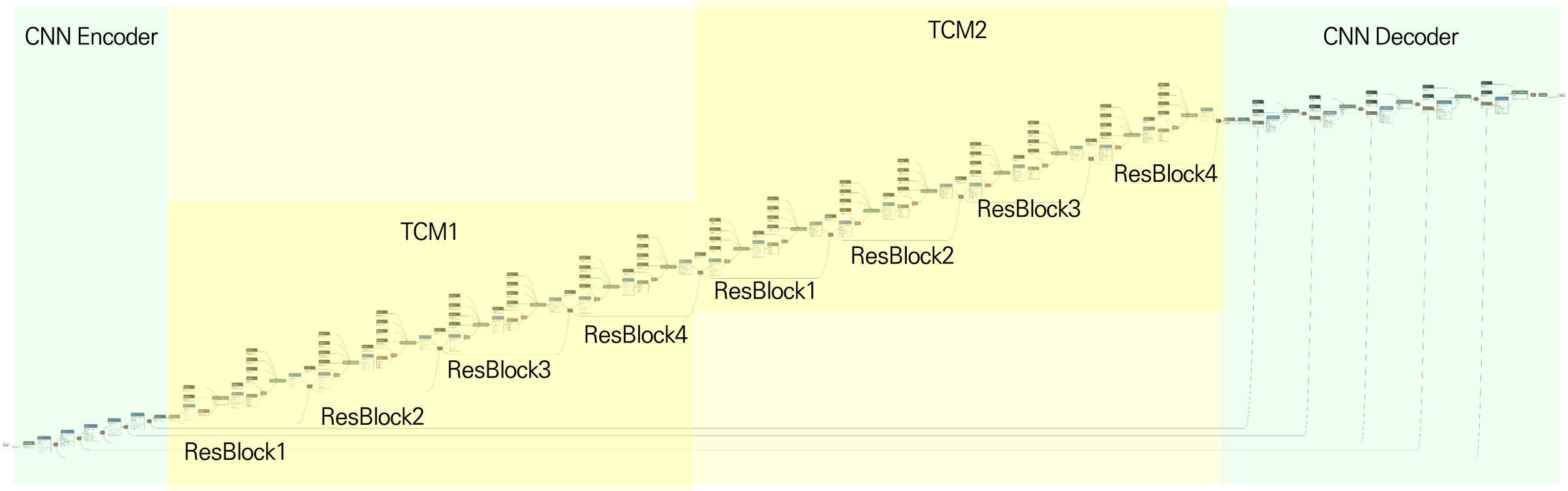
CNN+LSTM 모델 구조

➢ 전체 구조도



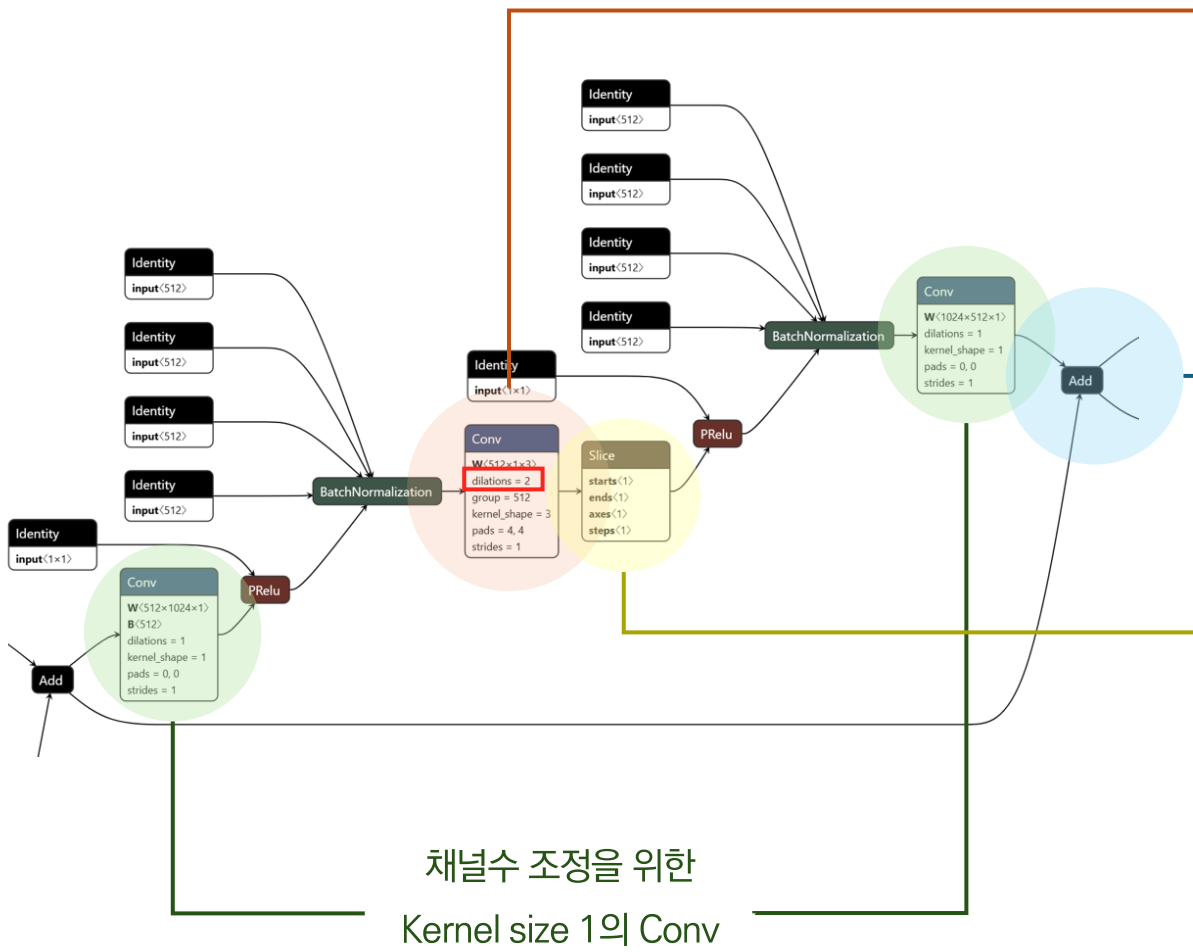
CNN+TCM 모델 구조

➢ 전체 구조도



CNN+TCM 모델 구조

➢ ResBlock 구조도



Dilated Conv

Dilation 값은 2^{i-1} (i번째 ResBlock에 대해)

Convolution을 거친 결과와 Input 값을
Add하여 Residual Learning

Causal한 시스템을 만들어주기 위해
미래의 정보가 고려된 부분을 제거

채널수 조정을 위한
Kernel size 1의 Conv

Inference time, RTF, STFT & iSTFT 소요시간 비교

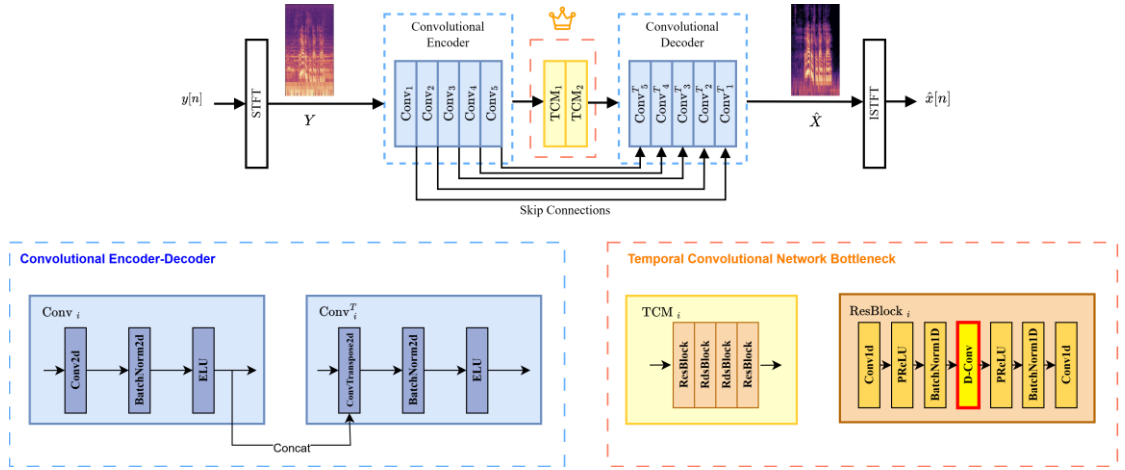
Sample Number	Audio 길이	Inference Time		RTF	
		Baseline (LSTM)	Proposed (TCM)	Baseline (LSTM)	Proposed (TCNN)
1	6.2s	2.00s	**0.47s**	0.321	**0.076**
2	4.7s	2.09s	**0.34s**	0.447	**0.073**
3	2.5s	0.84s	**0.21s**	0.332	**0.081**
4	3.8s	1.29s	**0.46s**	0.335	**0.120**

Sample Number	Audio 길이	STFT 소요시간		iSTFT 소요시간	
		Baseline (LSTM)	Proposed (TCM)	Baseline (LSTM)	Proposed (TCNN)
1	6.2s	0.013s	0.011s	0.024s	0.027
2	4.7s	0.003s	0.003s	0.007	0.005s
3	2.5s	0.003s	0.002s	0.005s	0.005s
4	3.8s	0.003s	0.003s	0.004s	0.006s

- LSTM을 사용한 Baseline 대비
  TCM을 사용한 Proposed의 Inference Time, RTF가 작은 값을 가짐

- 입력 waveform을 Spectrogram으로 변환하는 STFT 과정과
  Enhanced Spectrogram을 출력 waveform으로 복원하는 iSTFT 과정의
  소요시간은 모델 inference time에 비해 작은 값을 가짐

Reference TCNN vs Proposed TCNN



- Baseline En/Decoder + Reference TCNN
  - TCM 3개를 연결
  - TCM 내부 ResBlock 6개
  - 파라미터 수 19,733,991개

- Baseline En/Decoder + Proposed TCNN
  - TCM 2개를 연결
  - TCM 내부 ResBlock 4개
  - 파라미터 수 9,207,251개

Reference

- Tan, Ke, and DeLiang Wang. 2018. **"A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement."** *Interspeech 2018*, August. https://doi.org/10.21437/interspeech.2018-1405.

- Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. 2018. **"An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling."** arXiv.Org. March 4, 2018. https://arxiv.org/abs/1803.01271.

- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K. (2016) **WaveNet: A Generative Model for Raw Audio.** Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), 125

- A. Pandey and D. Wang, **"TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain,"** ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6875-6879