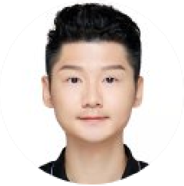
Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation

정승규

Paper information

BrAIn Lab.
Brain and Artificial Intelligence Lab.



Yi Luo

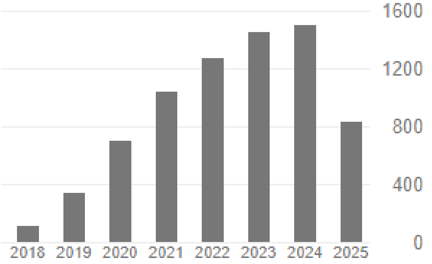
소속을 알 수 없음
columbia.edu의 이메일 확인됨
Speech separation speech enhancement

팔로우

제목	인용	연도
Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation Y Luo, N Mesgarani IEEE/ACM transactions on audio, speech, and language processing 27 (8), 1256 ...	2437	2019
Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation Y Luo, Z Chen, T Yoshioka ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and ...	987	2020
Tasnet: time-domain audio separation network for real-time, single-channel speech separation Y Luo, N Mesgarani 2018 IEEE International Conference on Acoustics, Speech and Signal ...	862	2018
Deep attractor network for single-microphone speaker separation Z Chen, Y Luo, N Mesgarani https://arxiv.org/abs/1611.08930	530	2016
Speaker-independent speech separation with deep attractor network Y Luo, Z Chen, N Mesgarani IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (4), 787-796	316	2018
Continuous speech separation: Dataset and analysis Z Chen, T Yoshioka, L Lu, T Zhou, Z Meng, Y Luo, J Wu, X Xiao, J Li ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and ...	263	2020

인용

	전체	2020년 이후
서지정보	7324	6827
h-index	25	25
i10-index	41	40



공개 액세스 모두 보기

자료 3개 자료 20개

비공개 공개

재정 지원 요구사항 기준

공동 저자

IEEE TASLP (IF:4.1)

Introduction

기존 speech separation 기법들은 time-frequency 도메인 기반

그러나 이 방법엔 단점 존재

1. STFT 는 음성 분리에 최적이라고 할 수 없다
2. 깨끗한 음원의 위상을 정확하게 복원하는것은 쉽지 않으며 잘못 복원된 위상은 정확도에 부정적인 영향
3. 시간-주파수 도메인에서 분리를 위해서는 혼합 신호의 고해상도 주파수 분해가 필요하여, 이는 STFT 계산을 위해 긴 시간 윈도우를 요구한다. -> 지연시간 증가

Introduction

이러한 문제를 해결하기 위해 1D CNN 기반 Unet 모델이 개발되었으나,

대규모 데이터에서 평가되지 않았음

TasNet 도 있으나, LSTM을 사용하여 여러가지 제약이 있음

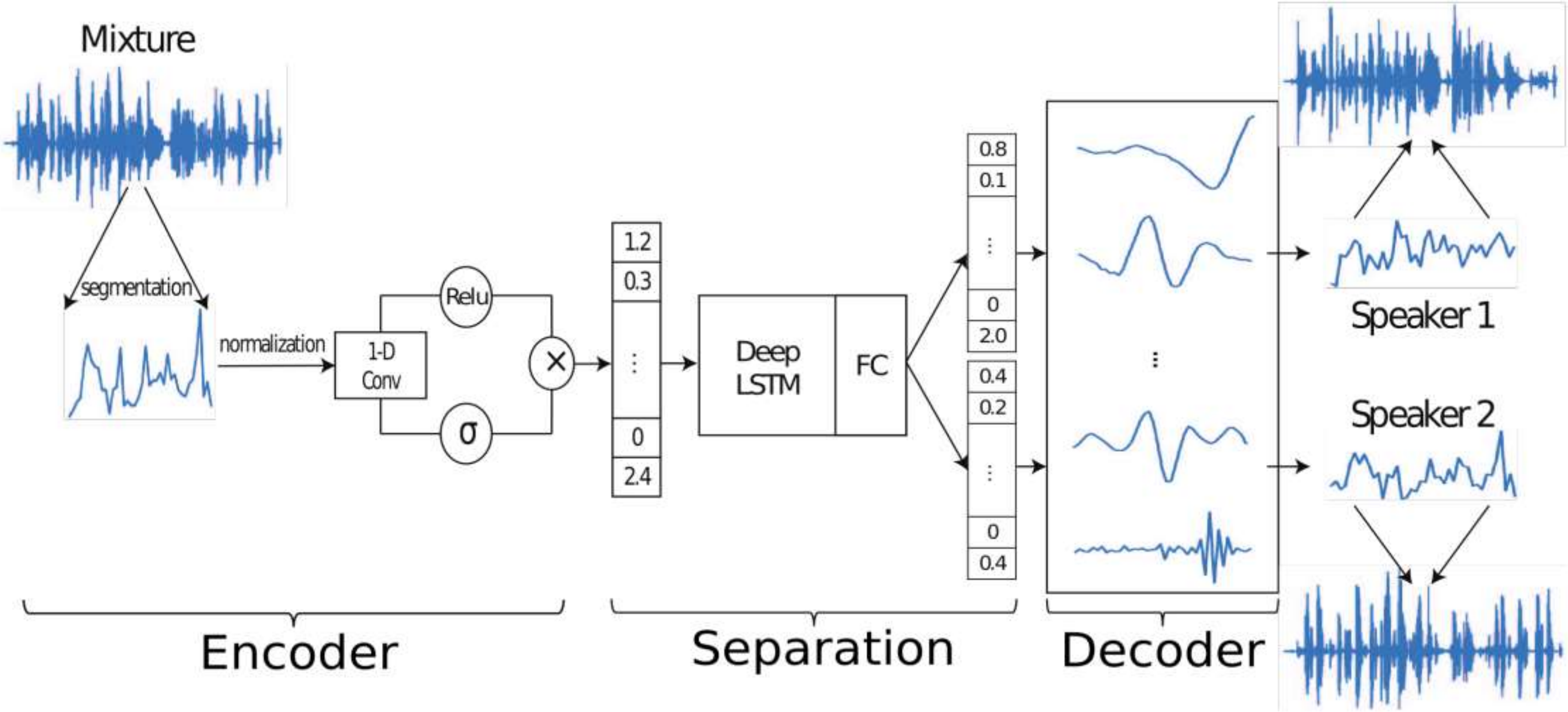
1. 인코더에서 더 작은 커널크기를 하면 LSTM 훈련 감당 안됨
2. LSTM 의 계산비용이 너무 큼

따라서 완전 컨볼루젼만 사용하는 Conv-TasNet 제안

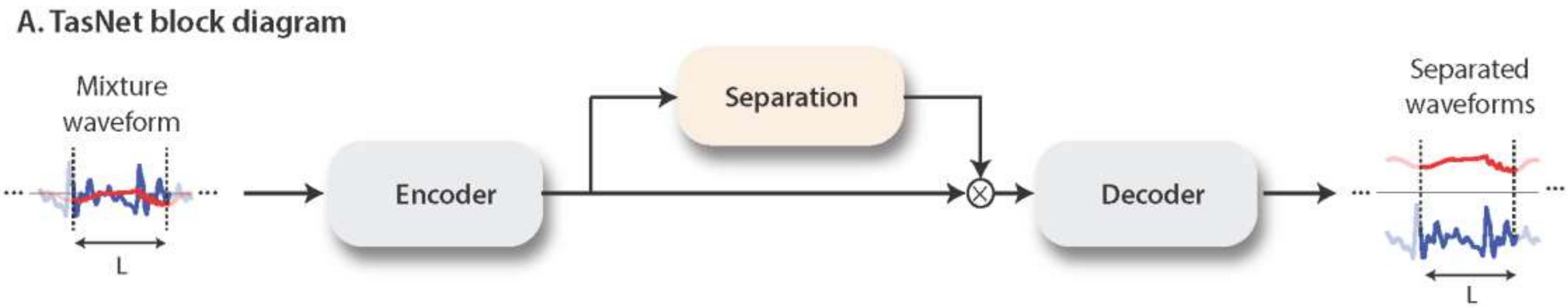
TCN 구조를 차용하여 dilated 1D Conv사용

연산량을 줄이기 위하여 depthwise separable conv 사용

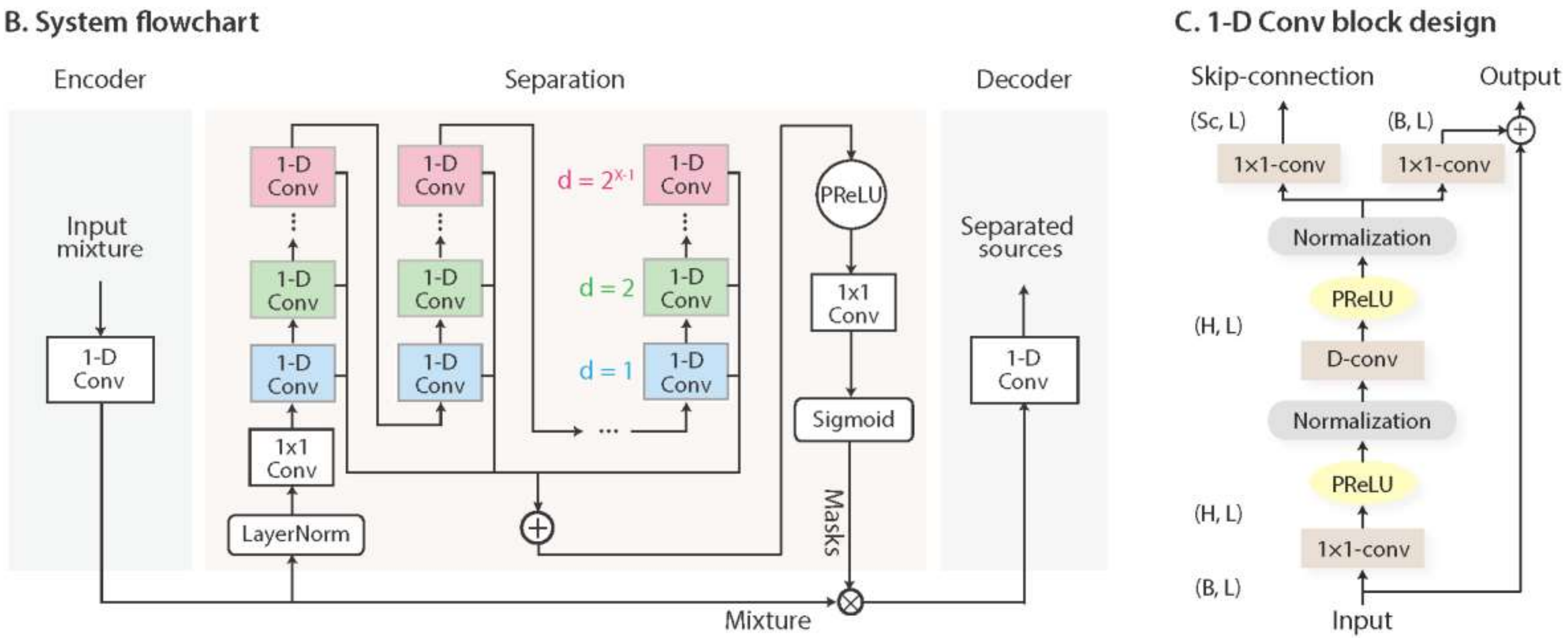
Introduction



Method


A. TasNet block diagram
Mixture
waveform
L
Encoder
Separation
Decoder
Separated
waveforms
L

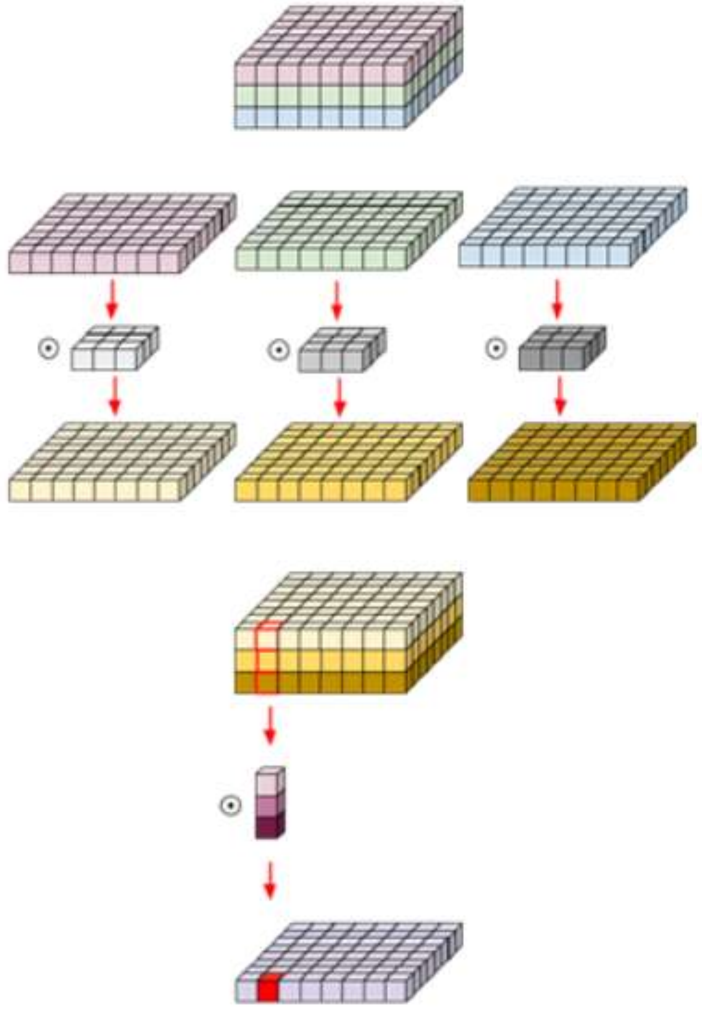
Method



B. System flowchart
Encoder
Separation
Decoder
Input mixture
1-D Conv
1x1 Conv
LayerNorm
d = 2^X-1
d = 2
d = 1
PReLU
Sigmoid
Masks
Mixture
Separated sources
C. 1-D Conv block design
Skip-connection
Output
(Sc, L)
(B, L)
1×1-conv
Normalization
PReLU
(H, L)
D-conv
1×1-conv
Input

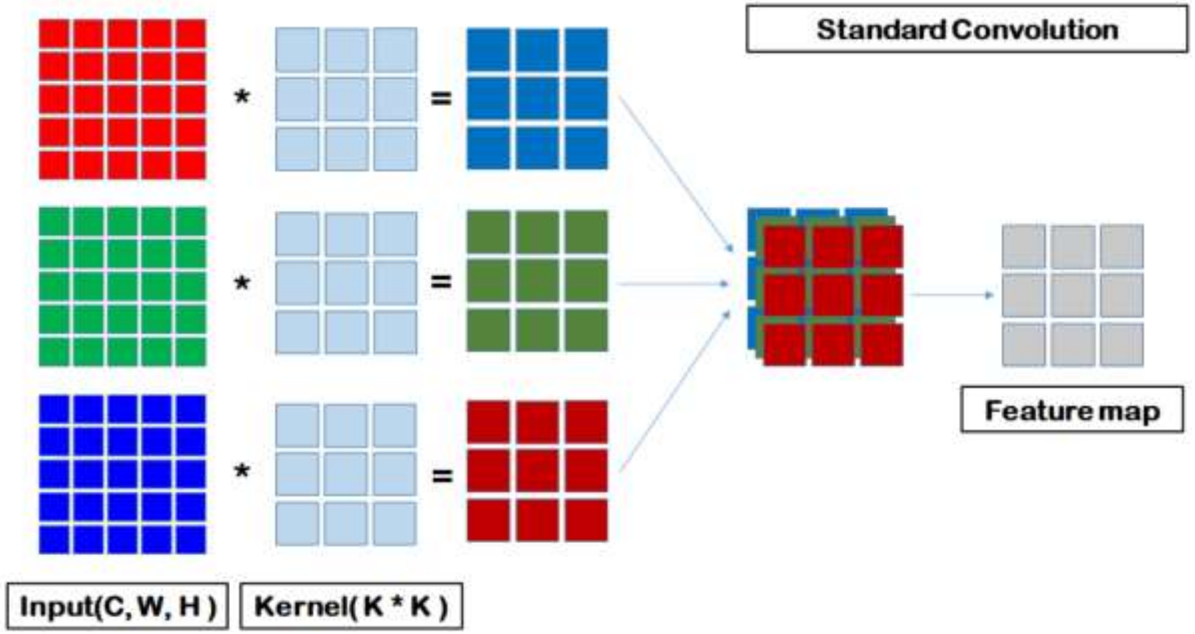
Method

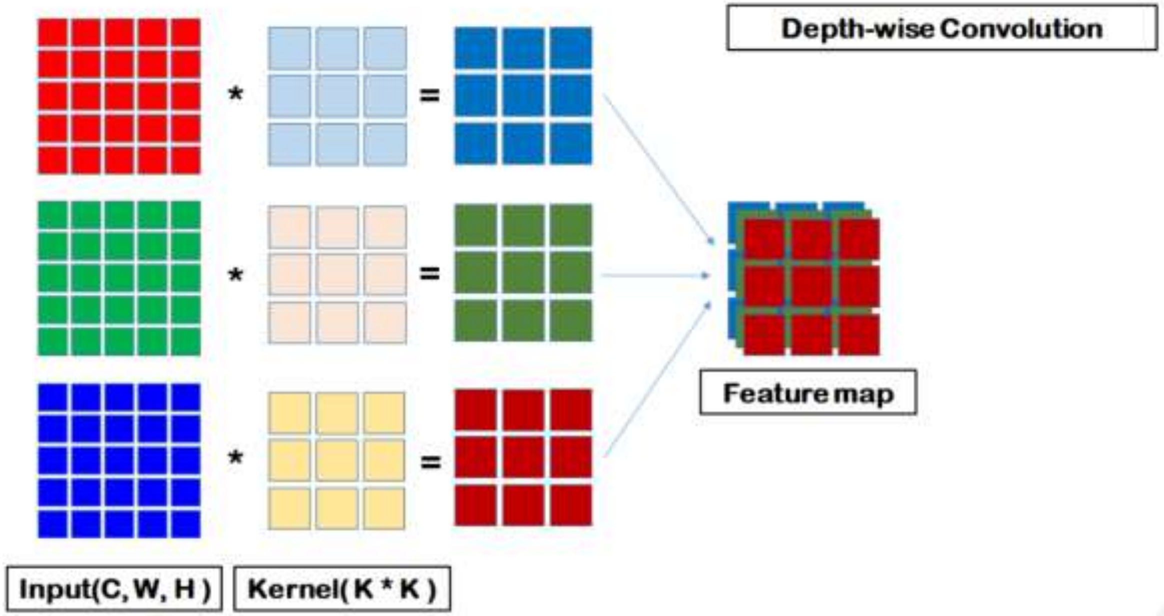
Depth-wise separable Convolution

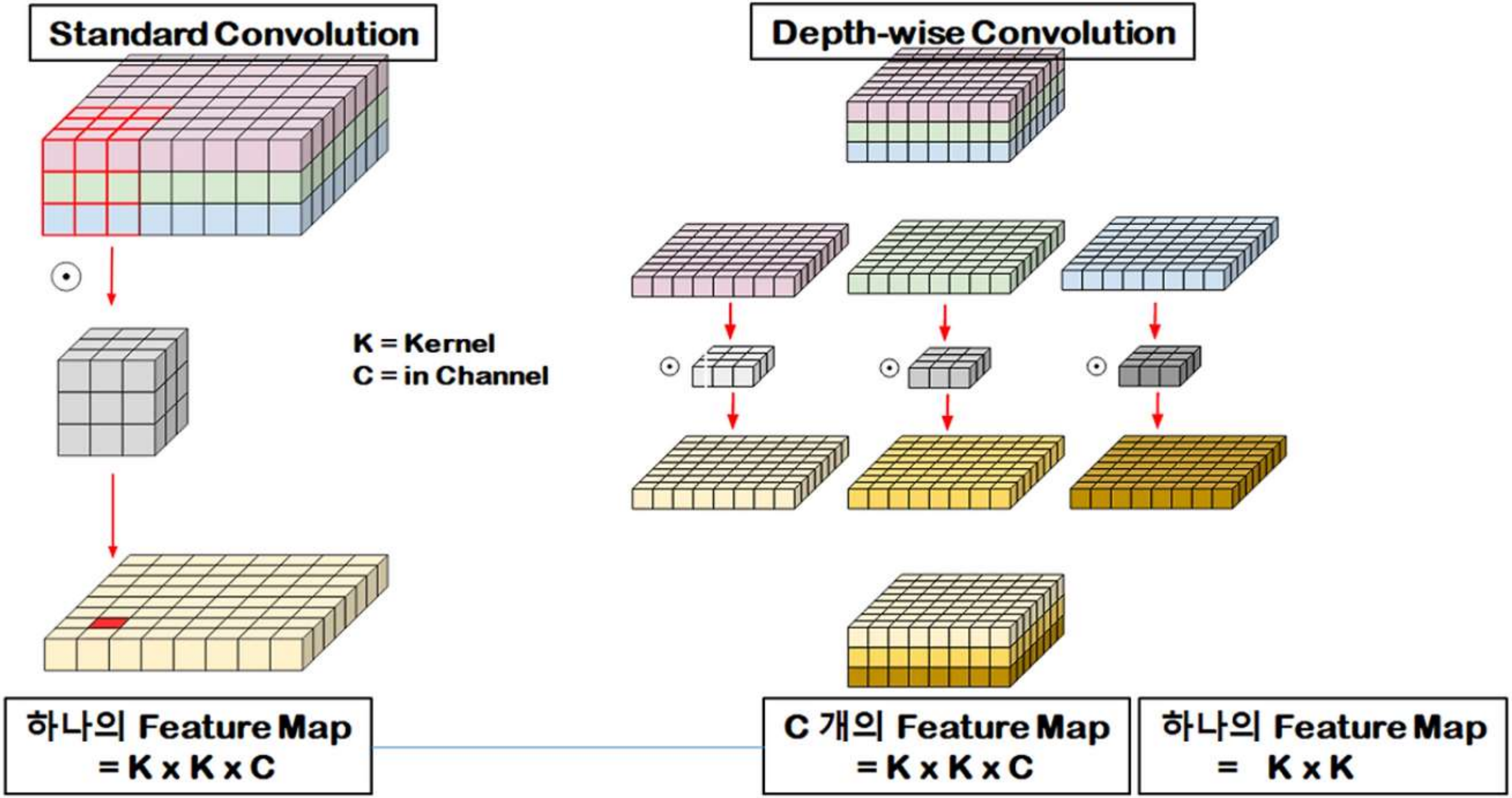


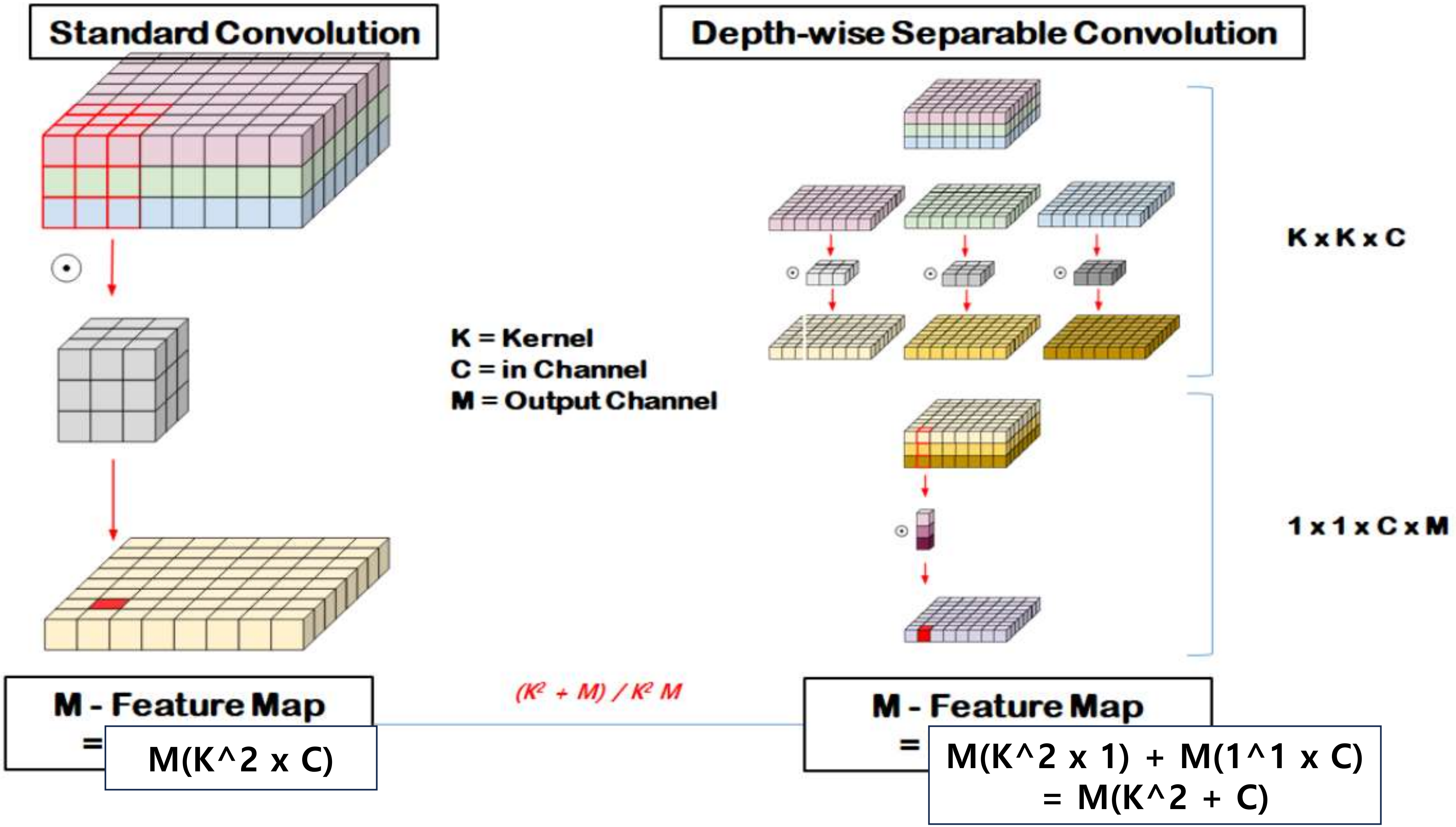
Method

Depth-wise Convolution

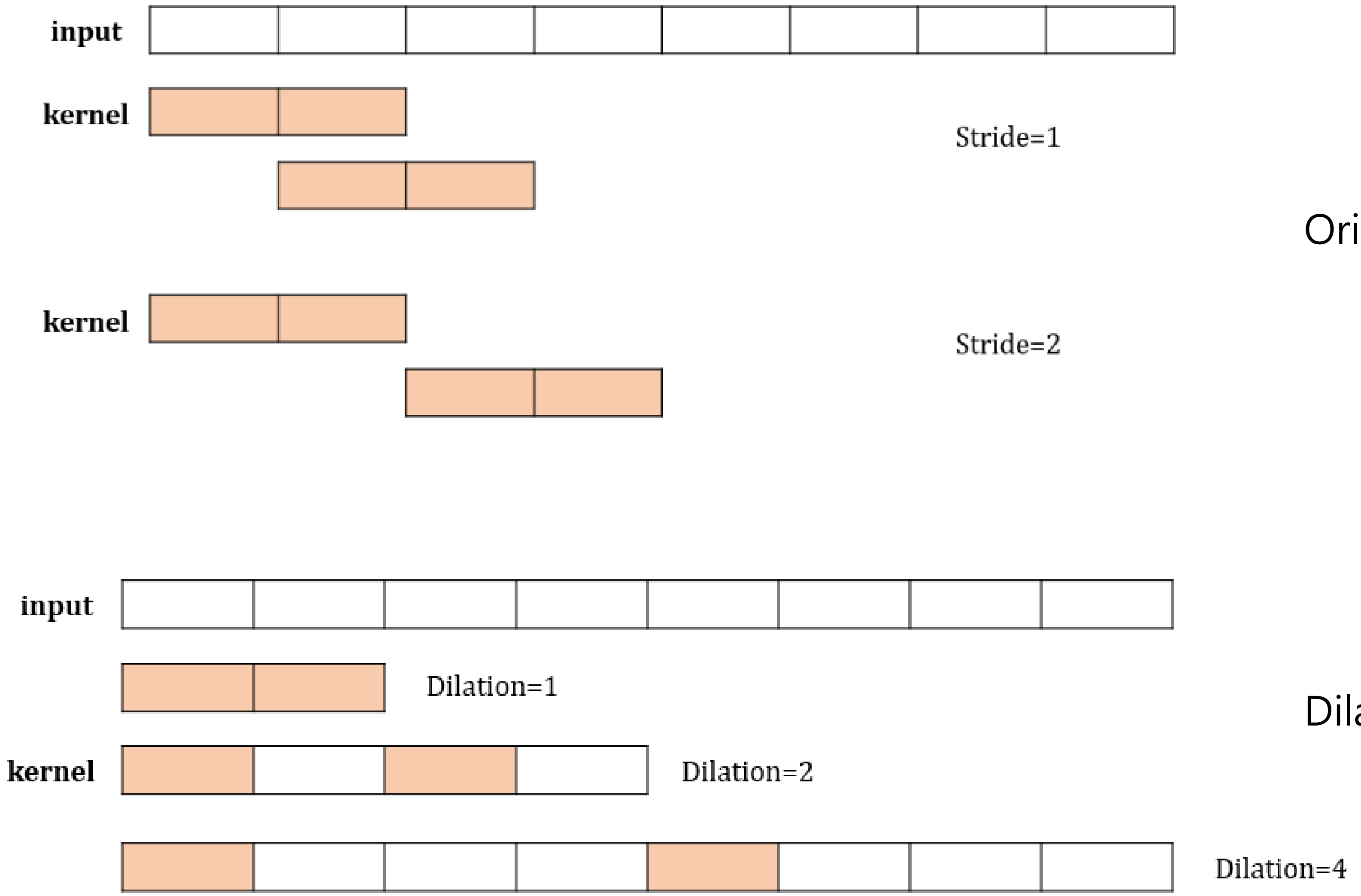





Standard Convolution
Depth-wise Convolution
K = Kernel
C = in Channel
하나의 Feature Map
= K x K x C
C 개의 Feature Map
= K x K x C
하나의 Feature Map
= K x K


Standard Convolution
Depth-wise Separable Convolution
K x K x C
K = Kernel
C = in Channel
M = Output Channel
1 x 1 x C x M
M - Feature Map
=
M(K^2 x C)
$(K^2 + M) / K^2 M$
M - Feature Map
=
M(K^2 x 1) + M(1^1 x C)
= M(K^2 + C)

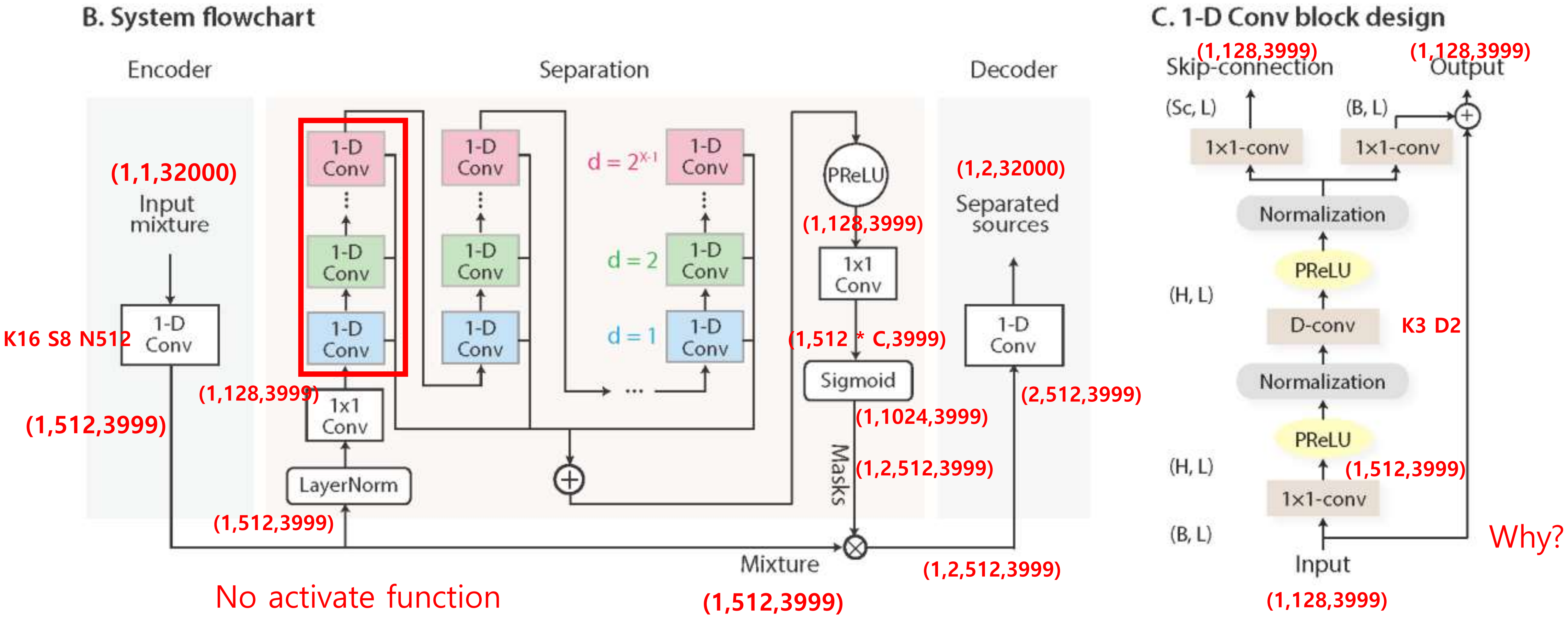
Method



Original

Dilation

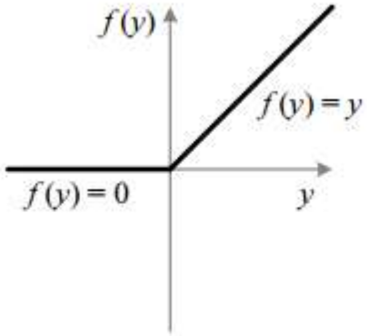
Method


B. System flowchart
Encoder
Separation
Decoder
(1,1,32000)
Input mixture
K16 S8 N512
1-D Conv
(1,512,3999)
1-D Conv
1-D Conv
1-D Conv
d = 2^(X-1)
d = 2
d = 1
(1,128,3999)
1x1 Conv
LayerNorm
(1,512,3999)
PReLU
(1,128,3999)
1x1 Conv
(1,512 * C,3999)
Sigmoid
(1,1024,3999)
Masks
(1,2,512,3999)
Mixture
(1,512,3999)
(1,2,512,3999)
(1,2,32000)
Separated sources
1-D Conv
(2,512,3999)
No activate function
C. 1-D Conv block design
(1,128,3999)
Skip-connection
(1,128,3999)
Output
(Sc, L)
(B, L)
1x1-conv
1x1-conv
Normalization
PReLU
(H, L)
D-conv
K3 D2
Normalization
PReLU
(H, L)
(1,512,3999)
1x1-conv
(B, L)
Input
(1,128,3999)
Why?

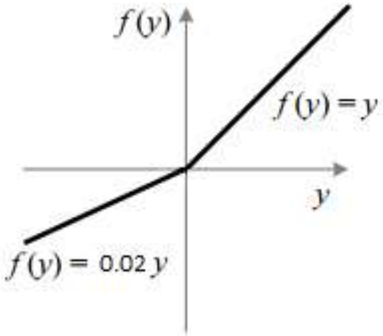
Method

PReLU

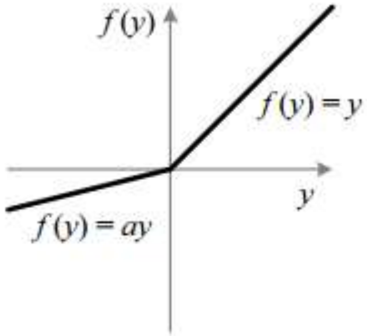
$$\mathrm{PReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{cases}$$



ReLU



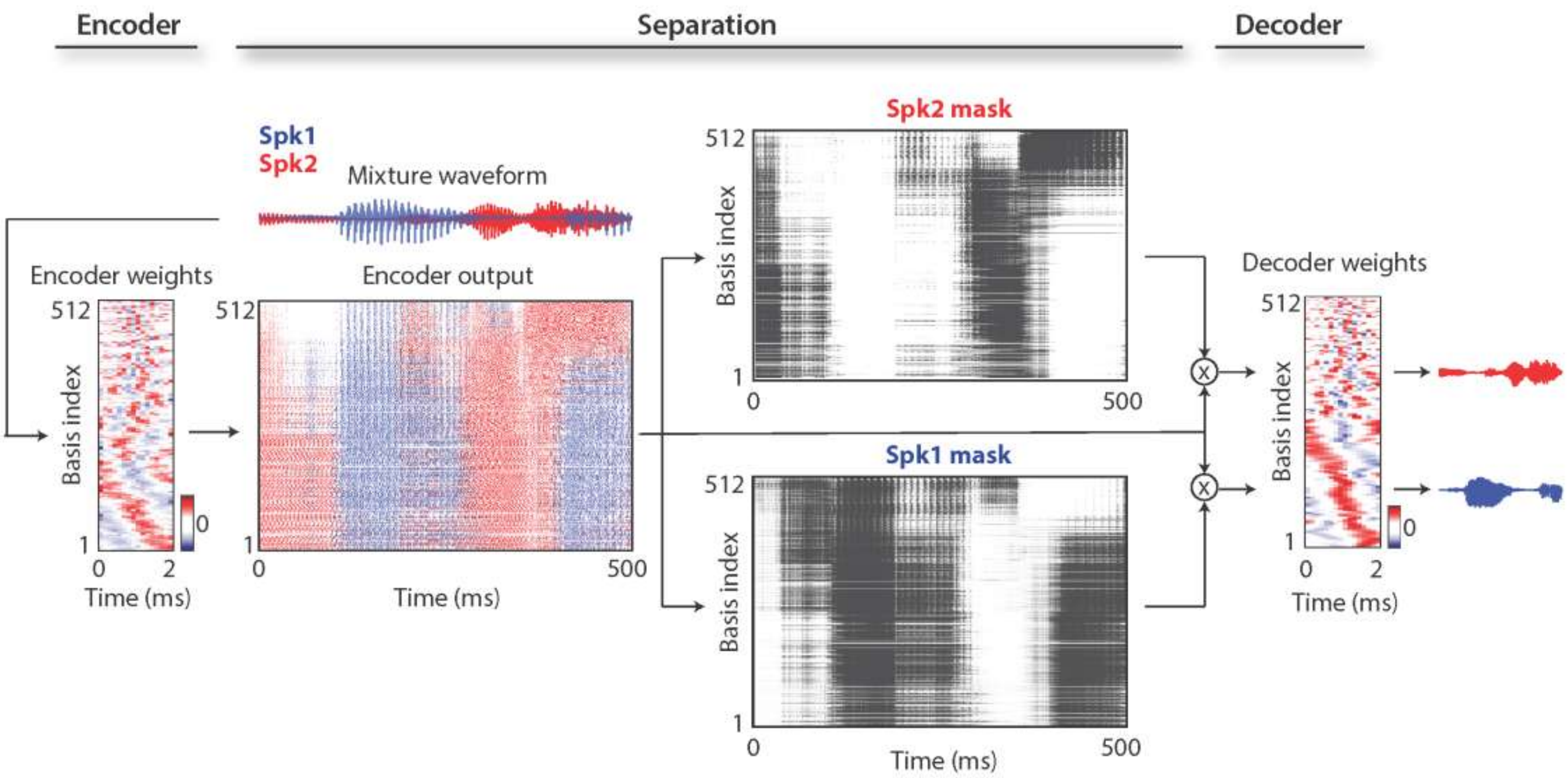
Leaky ReLU



PReLU

Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.



Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.

Dataset

WSJ0-2mix : 2 speech separation

- -5dB에서 +5dB 사이로 Mix
- Train : 30H
- Val : 10H
- Test 5H

WSJ0-3mix : 3 speech separation

Result

Eval Metric
Source-to-Distortion Ratio -> SDRi

$$\mathrm{SDR} = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{artif}\|^2}$$

Scale-Invariant Source-to-Noise Ratio -> SI-SNRi

$$s_{target} := \frac{\langle \hat{s}, s\rangle s}{\|s\|^2}$$

$$e_{noise} := \hat{s} - s_{target}$$

$$\text{SI-SNR} := 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{noise}\|^2}$$

Result

Symbol	Description
N	Number of filters in autoencoder
L	Length of the filters (in samples)
B	Number of channels in bottleneck and the residual paths' 1×1-*conv* blocks
Sc	Number of channels in skip-connection paths' 1×1-*conv* blocks
H	Number of channels in convolutional blocks
P	Kernel size in convolutional blocks
X	Number of convolutional blocks in each repeat
R	Number of repeats

TABLE II
THE EFFECT OF DIFFERENT CONFIGURATIONS IN CONV-TASNET.

N	L	B	H	Sc	P	X	R	Normalization	Causal	Receptive field (s)	Model size	SI-SNRi (dB)	SDRi (dB)
128	40	128	256	128	3	7	2	gLN	×	1.28	1.5M	13.0	13.3
256	40	128	256	128	3	7	2	gLN	×	1.28	1.5M	13.1	13.4
512	40	128	256	128	3	7	2	gLN	×	1.28	1.7M	13.3	13.6
512	40	128	256	256	3	7	2	gLN	×	1.28	2.4M	13.0	13.3
512	40	128	512	128	3	7	2	gLN	×	1.28	3.1M	13.3	13.6
512	40	128	512	512	3	7	2	gLN	×	1.28	6.2M	13.5	13.8
512	40	256	256	256	3	7	2	gLN	×	1.28	3.2M	13.0	13.3
512	40	256	512	256	3	7	2	gLN	×	1.28	6.0M	13.4	13.7
512	40	256	512	512	3	7	2	gLN	×	1.28	8.1M	13.2	13.5
512	40	128	512	128	3	6	4	gLN	×	1.27	5.1M	14.1	14.4
512	40	128	512	128	3	4	6	gLN	×	0.46	5.1M	13.9	14.2
512	40	128	512	128	3	8	3	gLN	×	3.83	5.1M	14.5	14.8
512	32	128	512	128	3	8	3	gLN	×	3.06	5.1M	14.7	15.0
512	16	128	512	128	3	8	3	gLN	×	1.53	5.1M	**15.3**	**15.6**
512	16	128	512	128	3	8	3	cLN	✓	1.53	5.1M	10.6	11.0

Result

BrAIn Lab.
Brain and Artificial Intelligence Lab.

TABLE III
SEPARATION SCORE FOR DIFFERENT SYSTEM CONFIGURATIONS.

Encoder	Mask	Model size	SI-SNRi (dB)	SDRi (dB)
Pinv	Softmax	1.5M	12.1	12.4
Linear	Softmax		12.9	13.2
	$\boxed{\text{Sigmoid}}$		**13.1**	**13.4**
ReLU	Softmax		13.0	13.3
	Sigmoid		12.9	13.2

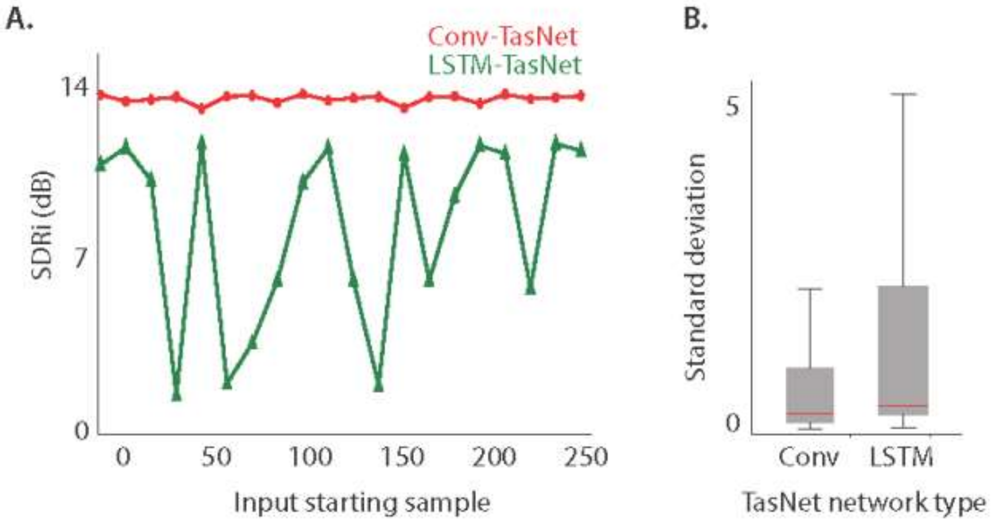
TABLE IV
COMPARISON WITH OTHER METHODS ON WSJ0-2MIX DATASET.

Method	Model size	Causal	SI-SNRi (dB)	SDRi (dB)
DPCL++ [5]	13.6M	×	10.8	–
uPIT-BLSTM-ST [7]	92.7M	×	–	10.0
DANet [8]	9.1M	×	10.5	–
ADANet [9]	9.1M	×	10.4	10.8
cuPIT-Grid-RD [50]	47.2M	×	–	10.2
CBLDNN-GAT [12]	39.5M	×	–	11.0
Chimera++ [10]	32.9M	×	11.5	12.0
WA-MISI-5 [11]	32.9M	×	12.6	13.1
BLSTM-TasNet [26]	23.6M	×	13.2	13.6
Conv-TasNet-gLN	**5.1M**	×	**15.3**	**15.6**
uPIT-LSTM [7]	46.3M	✓	–	7.0
LSTM-TasNet [26]	32.0M	✓	**10.8**	**11.2**
Conv-TasNet-cLN	**5.1M**	✓	10.6	11.0
IRM	–	–	12.2	12.6
IBM	–	–	13.0	13.5
WFM	–	–	13.4	13.8

Result

TABLE VIII
PROCESSING TIME FOR CAUSAL LSTM-TASNET AND CONV-TASNET. THE SPEED IS EVALUATED AS THE AVERAGE TIME REQUIRED TO SEPARATE A FRAME (TIME PER FRAME, TPF).

Method	CPU/GPU TPF (ms)
LSTM-TasNet	4.3/0.2
Conv-TasNet-cLN	**0.4/0.02**



Conclusion

핵심 기여

- 기존 TF 도메인이 아닌 Time 도메인으로 End to End 방법론
- 기존 Time 도메인 방법인 TasNet 대비 연산속도 대폭 축소 및 성능 향상

향후 과제 및 발전 방향

- 긴 대화 처리 어려움
- 잡음이나 울림이 심한 환경에서의 성능하락
- Multi channel 모델로 확장