Source Separation 1

안인규 (Inkyu An)

Speech And Audio Recognition (오디오 음성인식)

https://mairlab-km.github.io/

This lecture material refers to
https://github.com/yandexdataschool/speech_course?tab=readme-ov-file and
https://github.com/markovka17/dla

MAIR Lab

What is Source Separation?

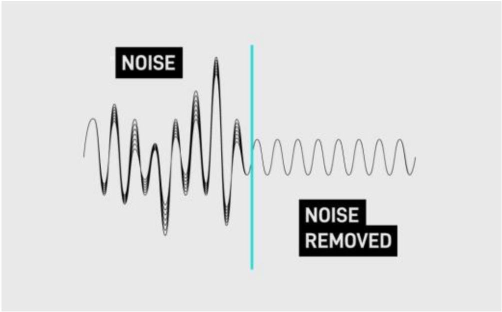
- Source Separation literally means separate any source of particular interest…

Separation?

- Separate signal (music, speech, …) from noise → **Denoising**
- Separate audio mix into separate tracks → **Source Separation**
- Guess when specific sound sources are active → **Diarization**

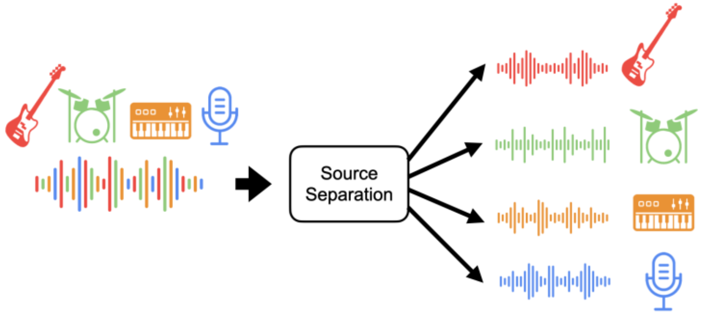
What is Source Separation?

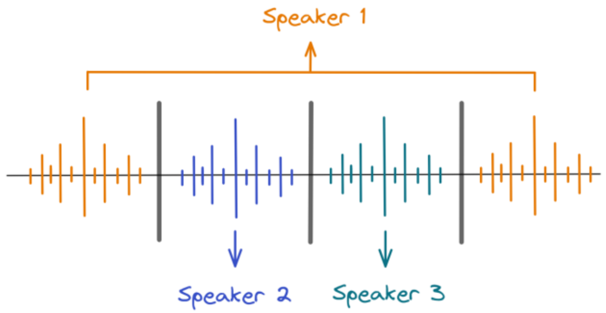
- Source Separation literally means separate any source of particular interest...



Separation?

- Separate signal (music, speech, ...) from noise → **Denoising**
- Separate audio mix into separate tracks → **Source Separation**
- Guess when specific sound sources are active → **Diarization**





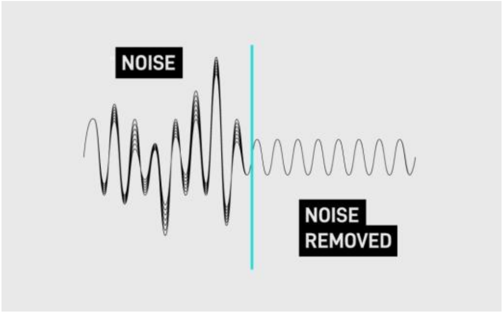
What is Source Separation?

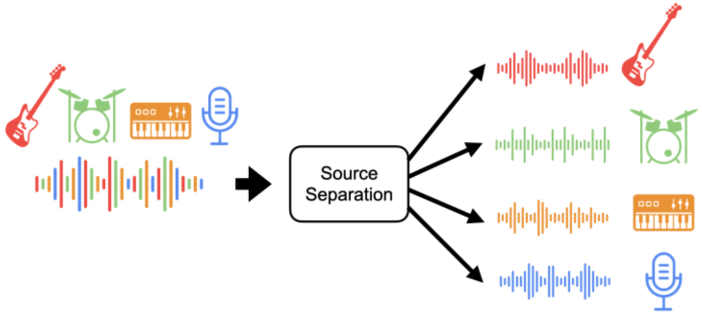
- Source Separation literally means separate any source of particular interest…

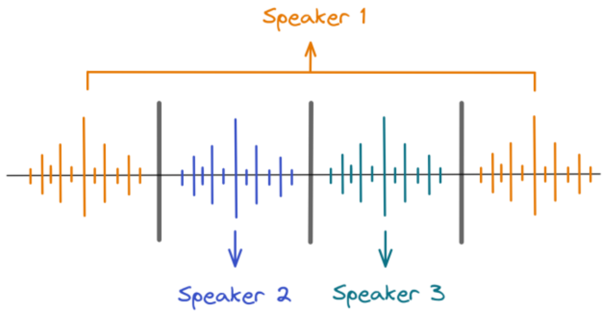
Source Separation + Denoising

Separation?

- Separate signal (music, speech, …) from noise → **Denoising**
- Separate audio mix into separate tracks → **Source Separation**
- Guess when specific sound sources are active → **Diarization**







Denoising Problem

- **Goal**: guess what is noise and remove the noise from audio

Sound is Simple

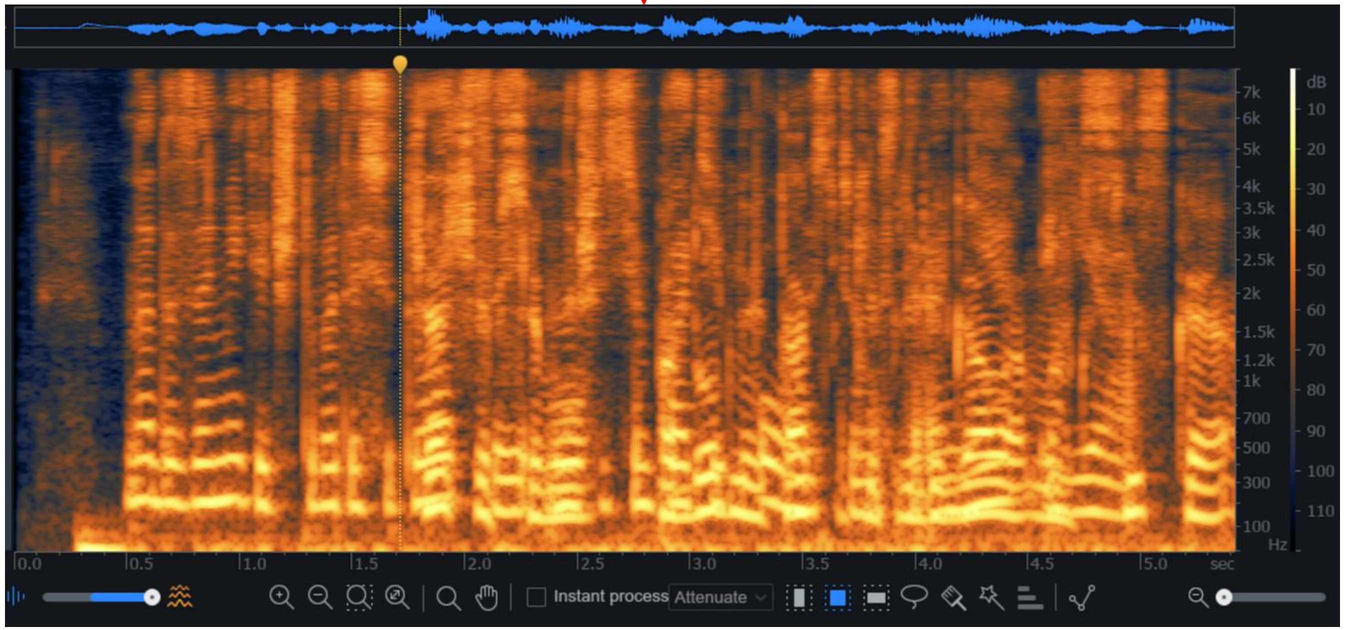
- **Spectrum-based Denoising**
  - **Main idea**: <u>the noise is different</u>, we will just cut it from the spectrum
  - **Early (and still popular) solution**: different types of spectral profiling

Ref.: Izotope RX 8

Denoising Problem

- **Goal**: guess what is noise and remove the noise from audio

Sound is complicated (Two people speaking)



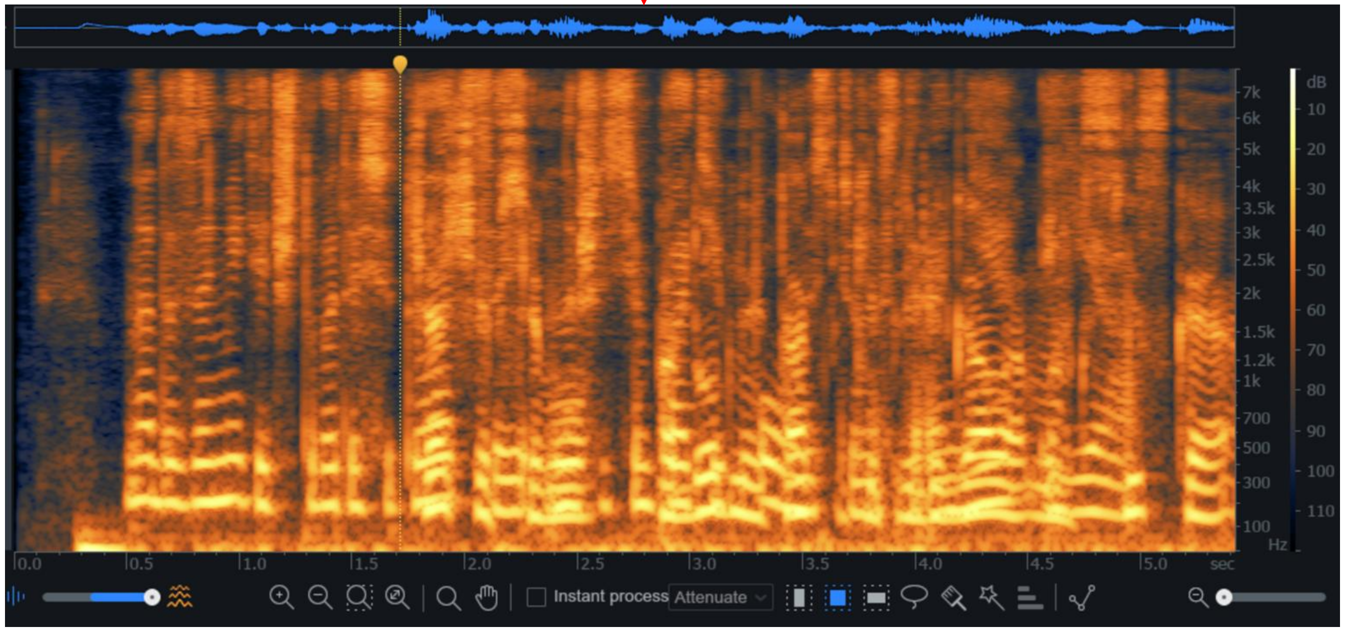
- **Spectrum-based Denoising**
  - **Main idea**: <u>the noise is different</u>, we will just cut it from the spectrum
  - **Early (and still popular) solution**: different types of spectral profiling

Ref.: Izotope RX 8

Denoising Problem

- **Goal**: guess what is noise and remove the noise from audio

Sound is complicated (Two people speaking) → The noise may not be well-separable



- **Spectrum-based Denoising**
  - **Main idea**: the noise is different, we will just cut it from the spectrum
  - **Early (and still popular) solution**: different types of spectral profiling

Ref.: Izotope RX 8

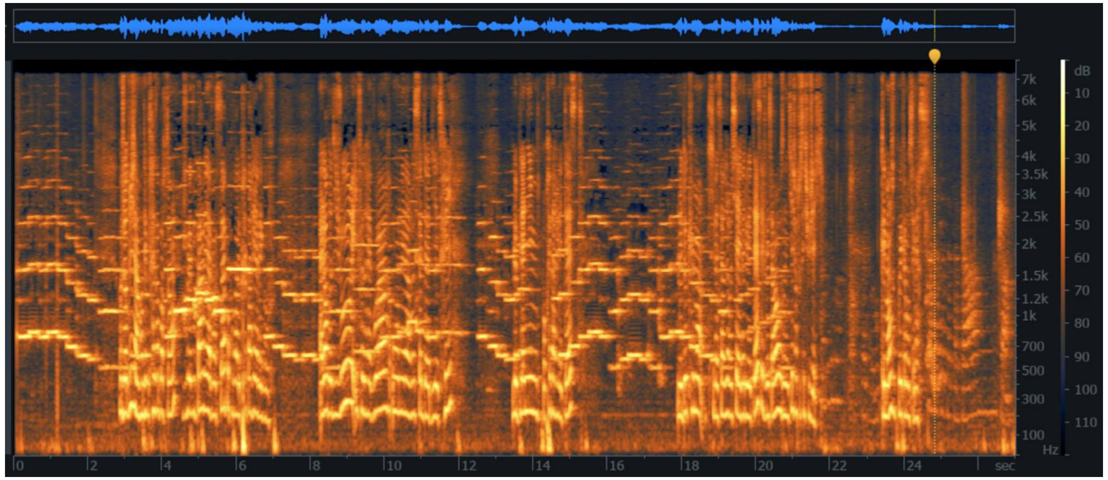
Denoising Problem

- **Deep learning approach**: see what can be done on the spectrogram

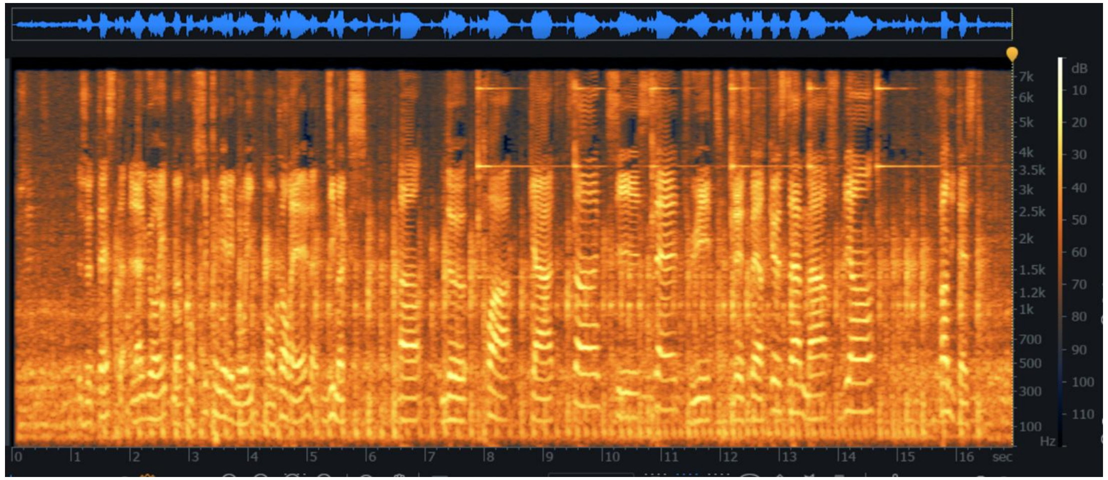
- **Main idea**: we'll still just cut it from the "spectrum"

- To separate complex audio, we need nontrivial ways

Sarah & Flute



Alex & noise

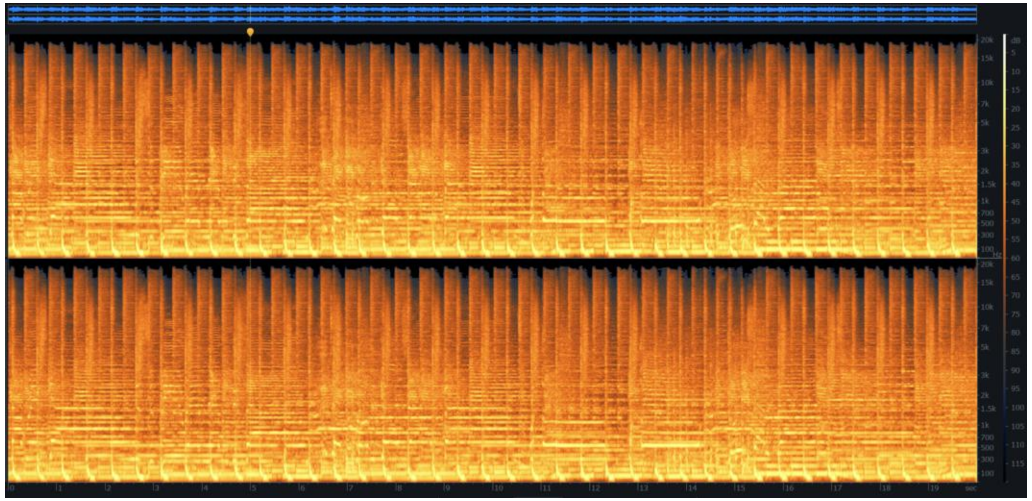


Audio Ref: DEMUCS

Separation Problem

- **Deep learning approach**: very similar

Song

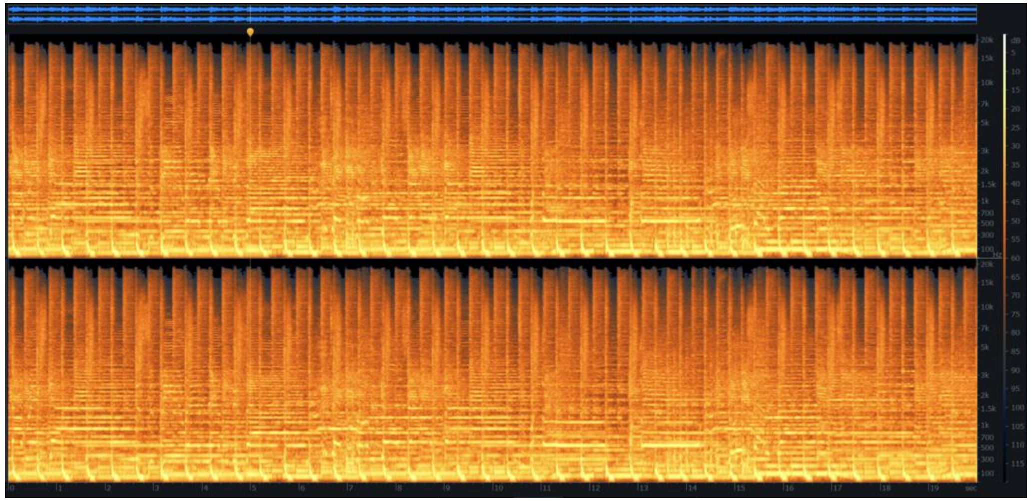


Audio Ref: [DEMUCS]

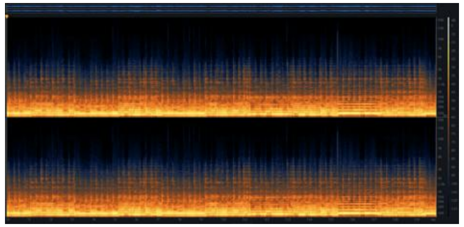
Separation Problem

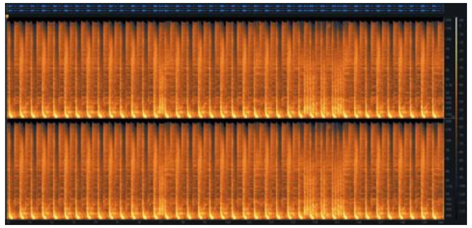
- **Deep learning approach**: very similar
- **Main idea**: still just carve it from the "spectrum" (using DEMUCS)

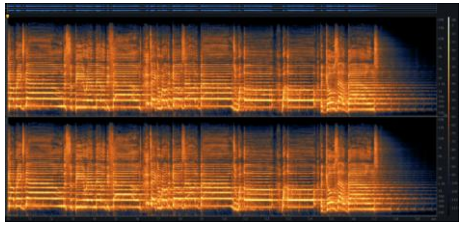
Song



The Easton Ellises - Falcon 69







Bass:

Drums:

Vocals:

True:

True:

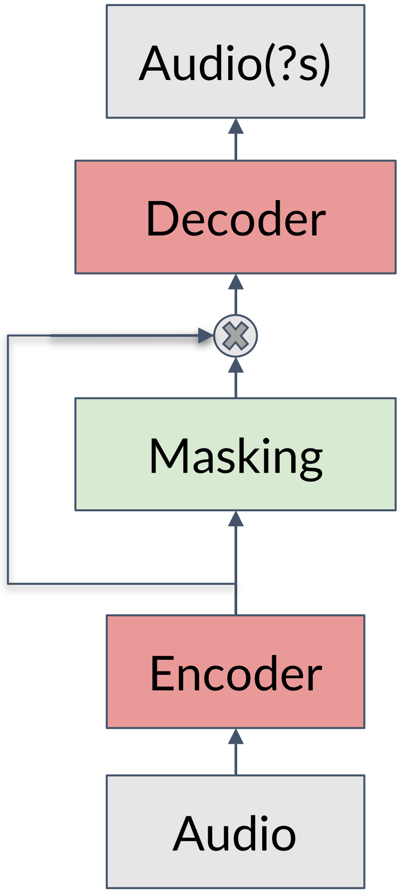
True:

Audio Ref: [DEMUCS]

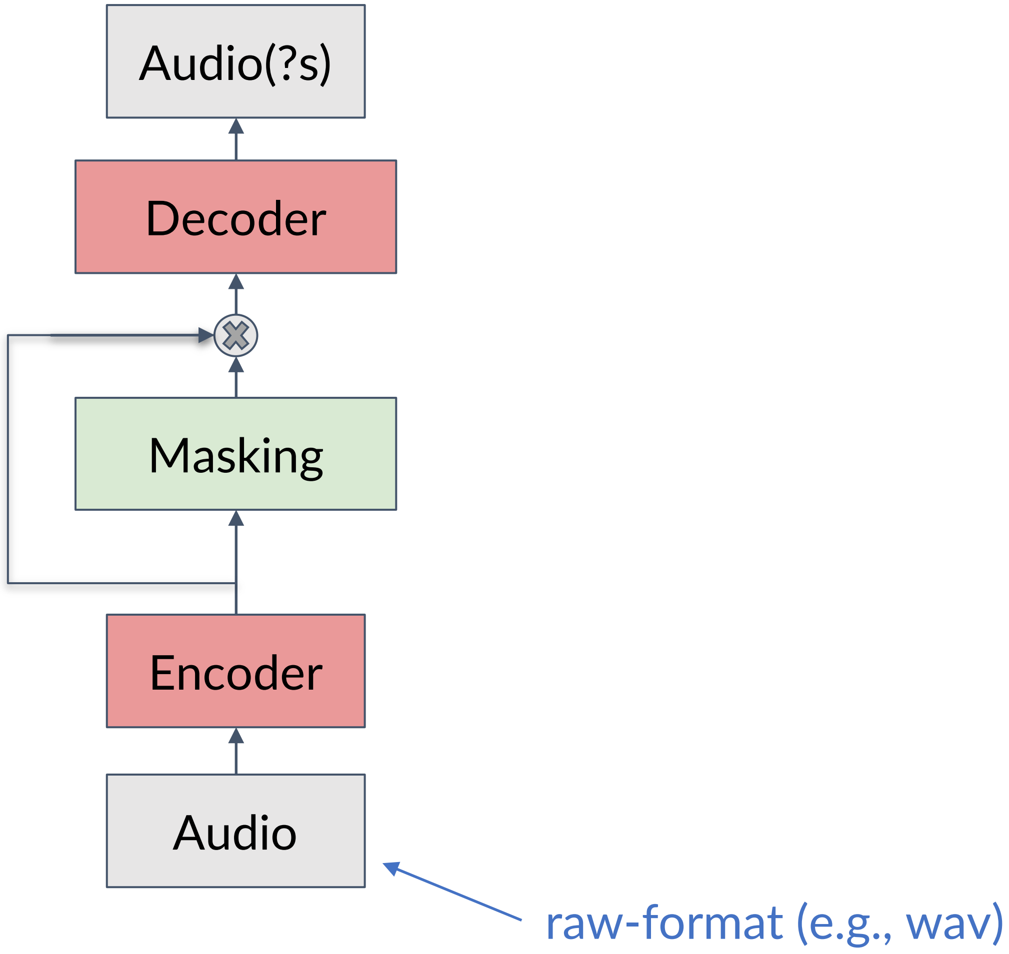
Separation Applications

- …

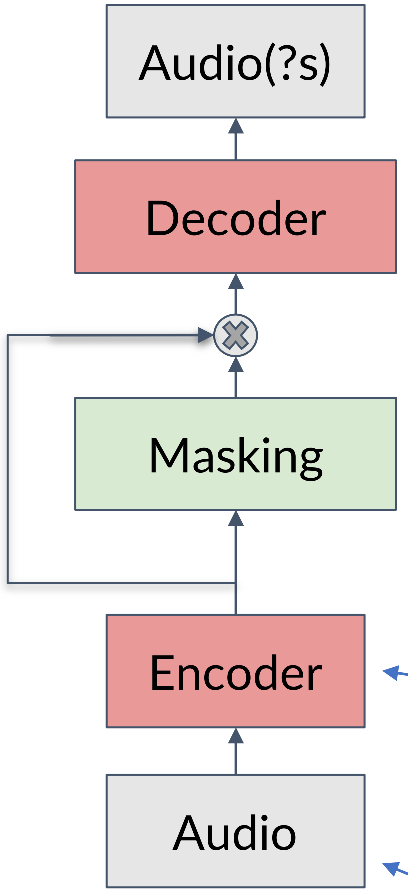
Denoising and Separation Pipeline



Denoising and Separation Pipeline



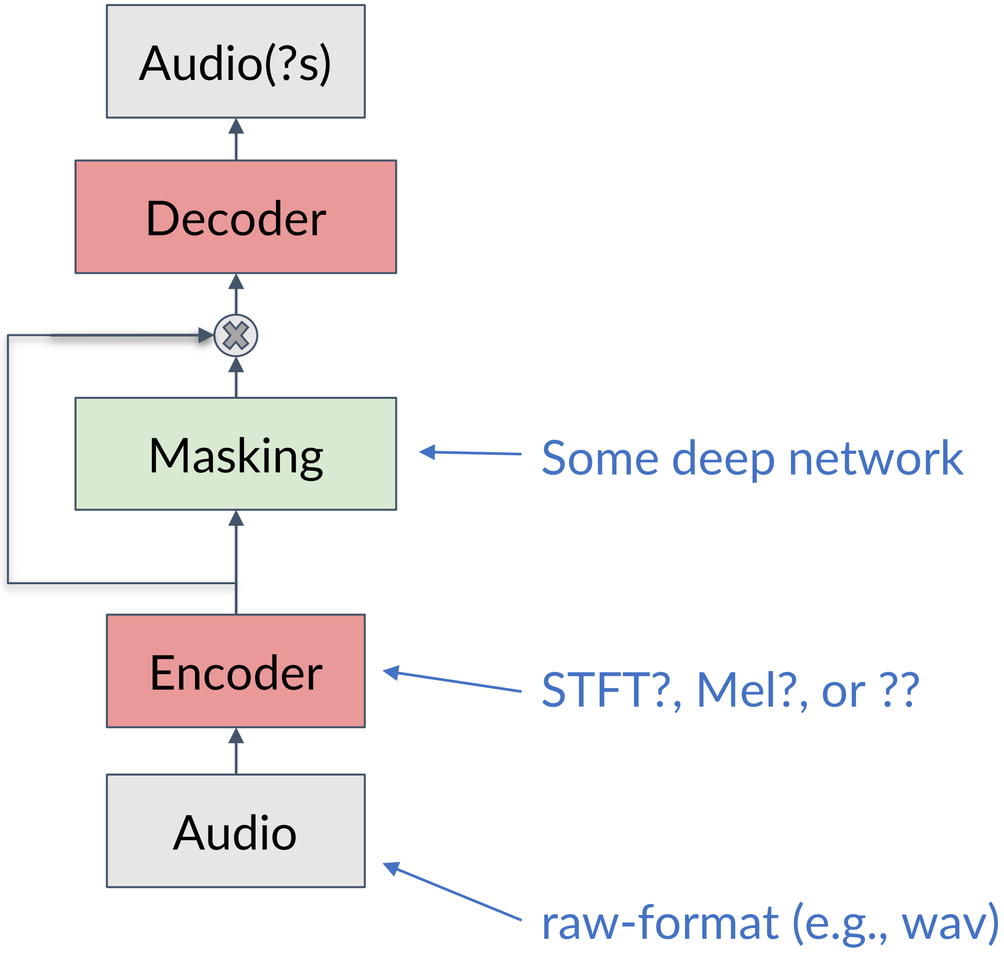
Denoising and Separation Pipeline



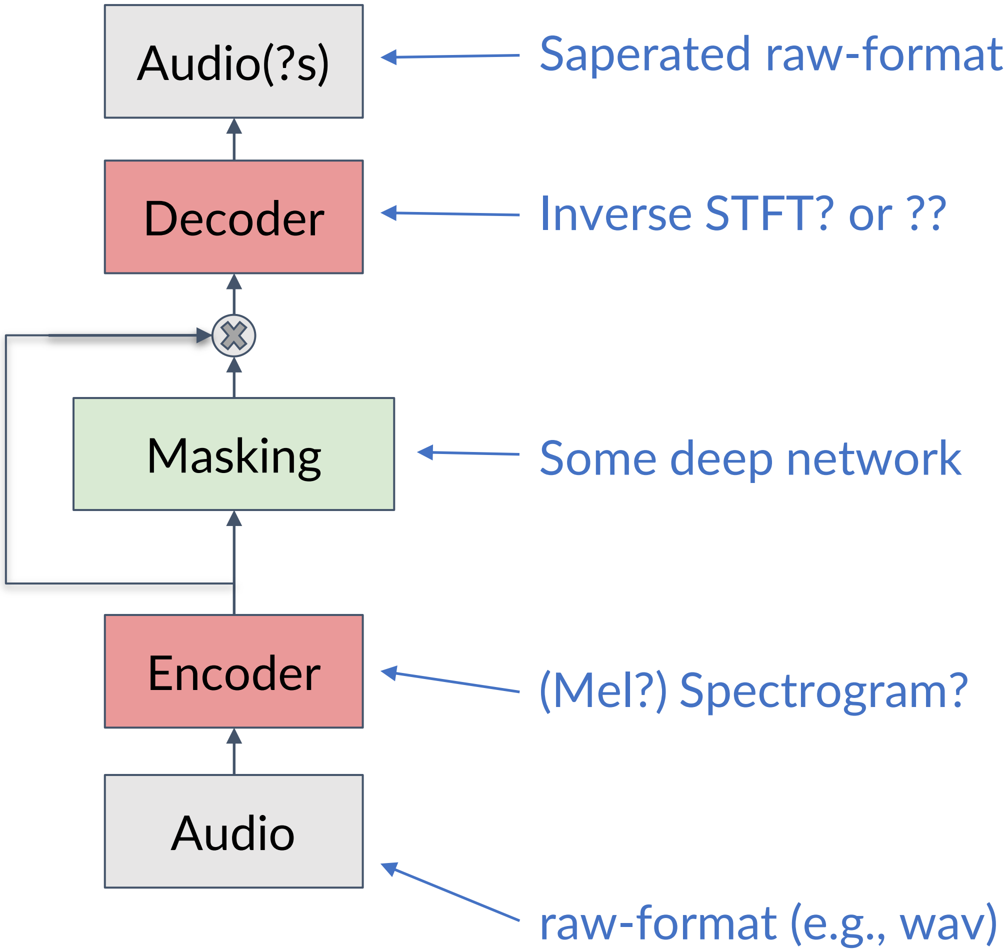
STFT?, Mel?, or ??

raw-format (e.g., wav)

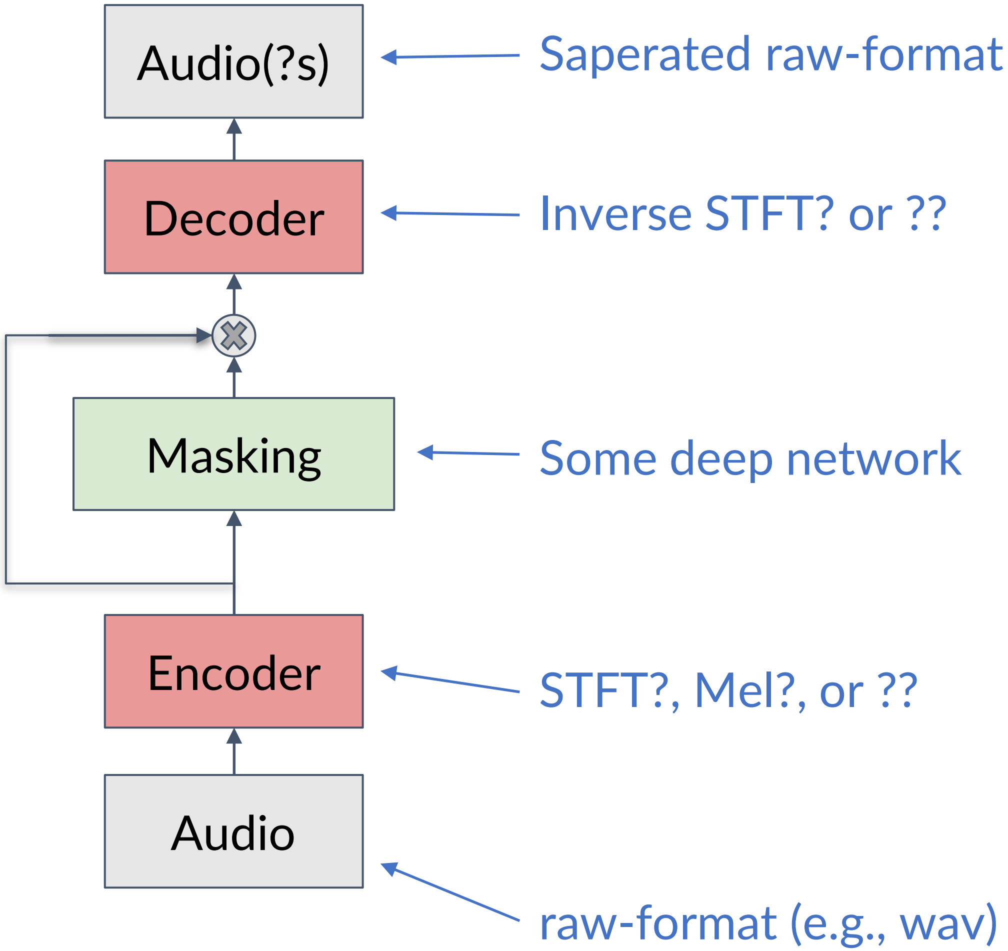
Denoising and Separation Pipeline



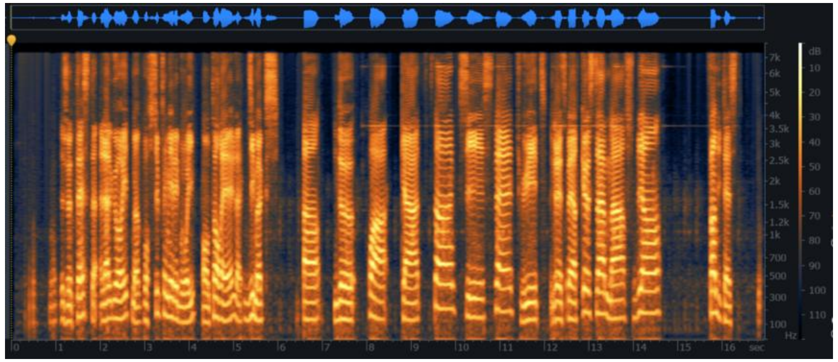
Denoising and Separation Pipeline



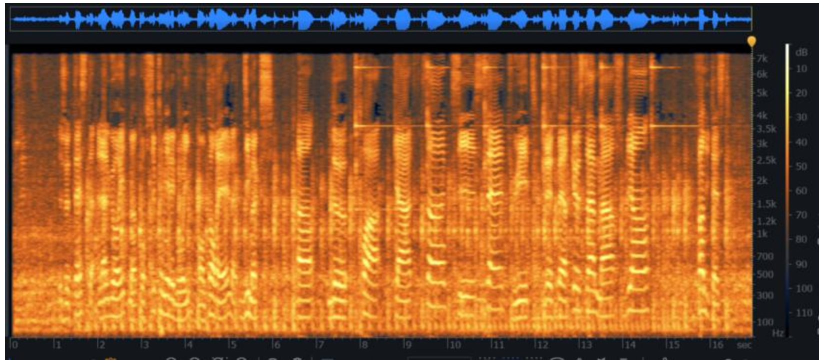
Denoising and Separation Pipeline



Denoised
Alex



Alex
&
noise

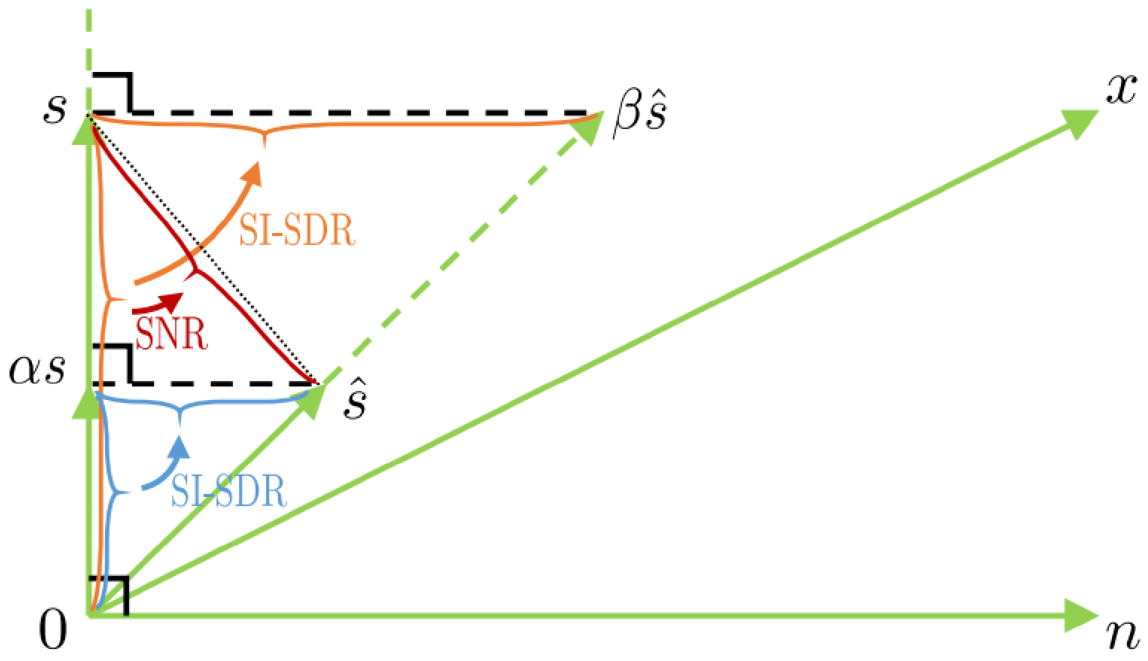


Denoising and Separation Metrics

- SNR (Signal-to-Noise Ratio) in dB:
- Si-SNR (Scale-invariant SNR) in dB:
- PESQ (Perceptual Evaluation of Speech Quality):
- STOI (Short-Time Objective Intelligibility):

Denoising and Separation Metrics

- SNR (Signal-to-Noise Ratio) and Si-SNR (Scale-invariant SNR)



Clean (s)

Separated (s+n)

$$SNR = 10\log\frac{\|s\|^2}{\|s-\hat{s}\|^2}$$

$$SI\_SNR = 10\log\frac{\|\alpha s\|^2}{\|\alpha s-\hat{s}\|^2}$$

SI-SNR이 언제 효과적일까?

https://ar5iv.labs.arxiv.org/html/1811.02508

Denoising and Separation Metrics

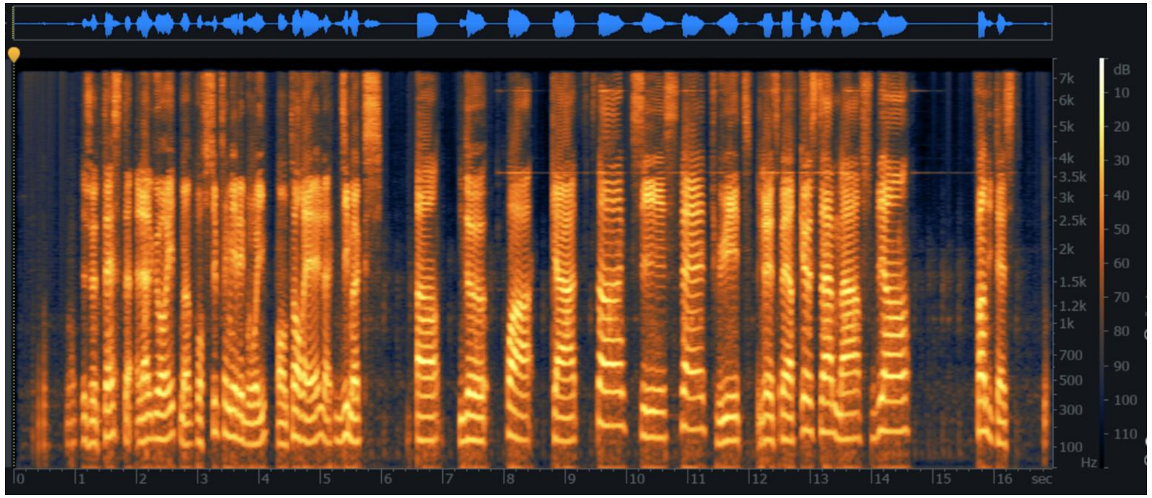
- PESQ (Perceptual Evaluation of Speech Quality):
  - 사람이 느끼는 음질 (Perceptual quality)를 객관적으로 수치화하기 위한 지표
  - -0.5 (bad)~4.5 (great)

- STOI (Short-Time Objective Intelligibility)
  - 사람이 얼마나 말을 알아들을 수 있는가 (intelligibility)를 예측하기 위한 지표
  - 0 (Bad)~1 (great)

https://ar5iv.labs.arxiv.org/html/1811.02508

Spectrogram Information

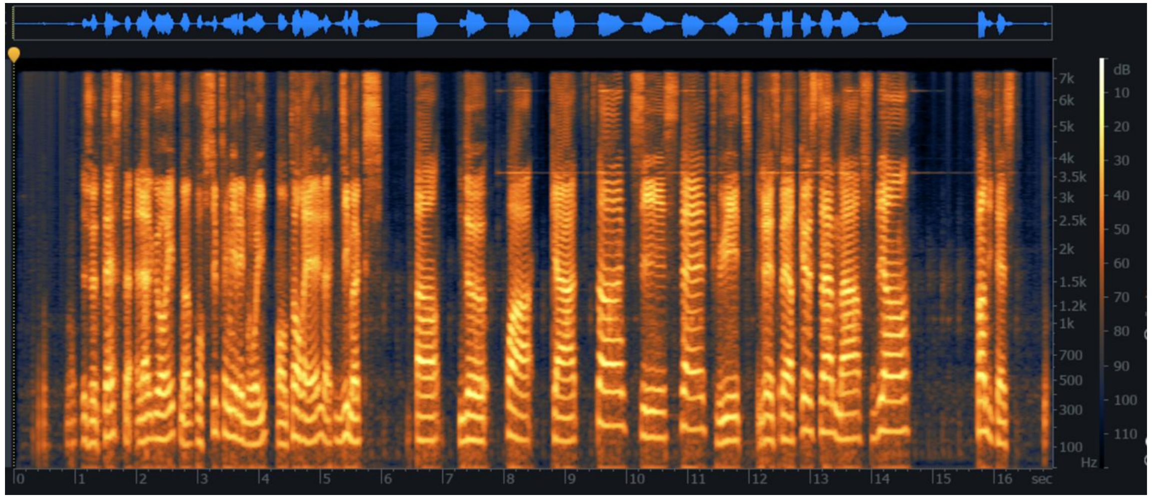
- **Spectrogram** (<u>Amplitude info.</u>)
  - Convert the magnitude (or squared magnitude) of STFT to a dB scale



Spectrogram Information

- **Spectrogram** (Amplitude info.)
    - Convert the magnitude (or squared magnitude) of STFT to a dB scale
    - **Idea**: use spectrogram as part of the encoder and inverse STFT as a decoder from masked spectrogram

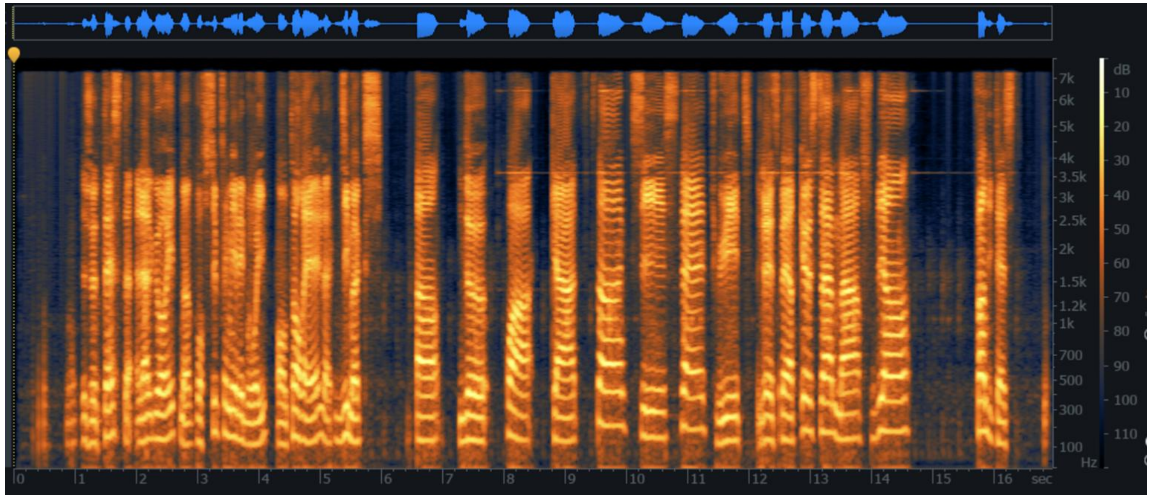
    - How to compute loss?
    - How to encode Spectrogram?



Spectrogram Information

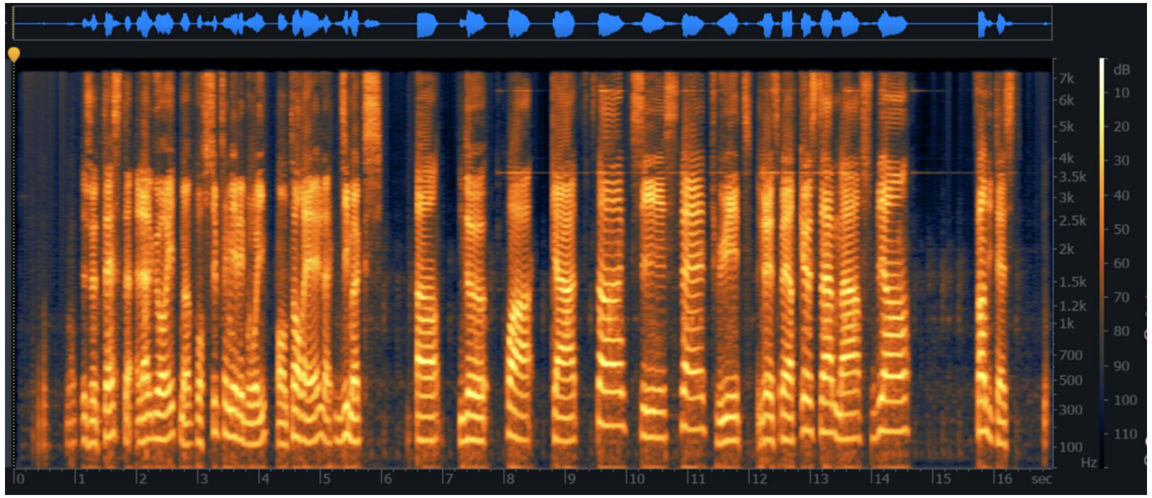
- **Spectrogram** (<u>Amplitude info.</u>)
  - Convert the magnitude (or squared magnitude) of STFT to a dB scale
  - **Idea**: use spectrogram as part of the encoder and inverse STFT as a decoder from masked spectrogram

  - How to compute loss?
  - How to encode Spectrogram?
  - How to handle Phase info.?

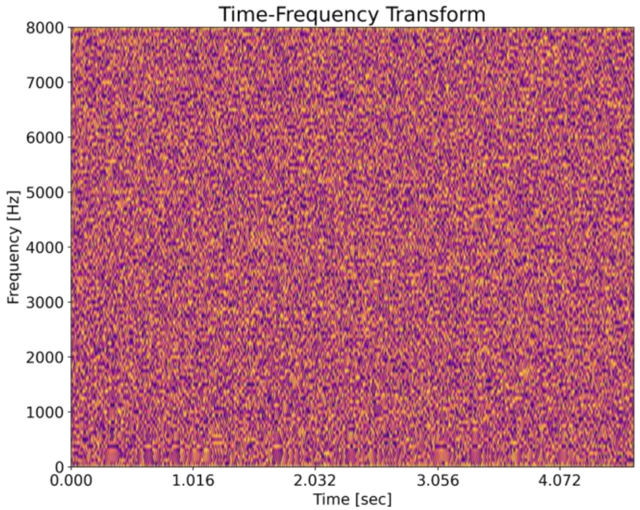


Spectrogram Information

- **Spectrogram** (<u>Amplitude info.</u>)



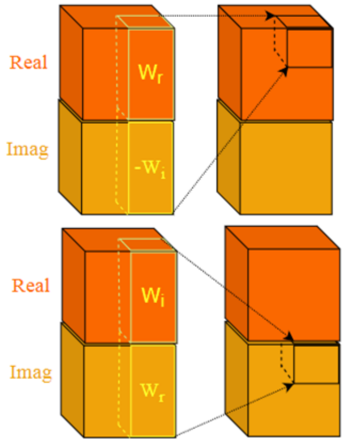
+



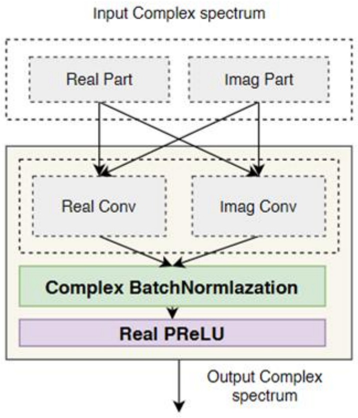
How can we use phase effectively?

DCCRN (DNS Challenge, 2020)

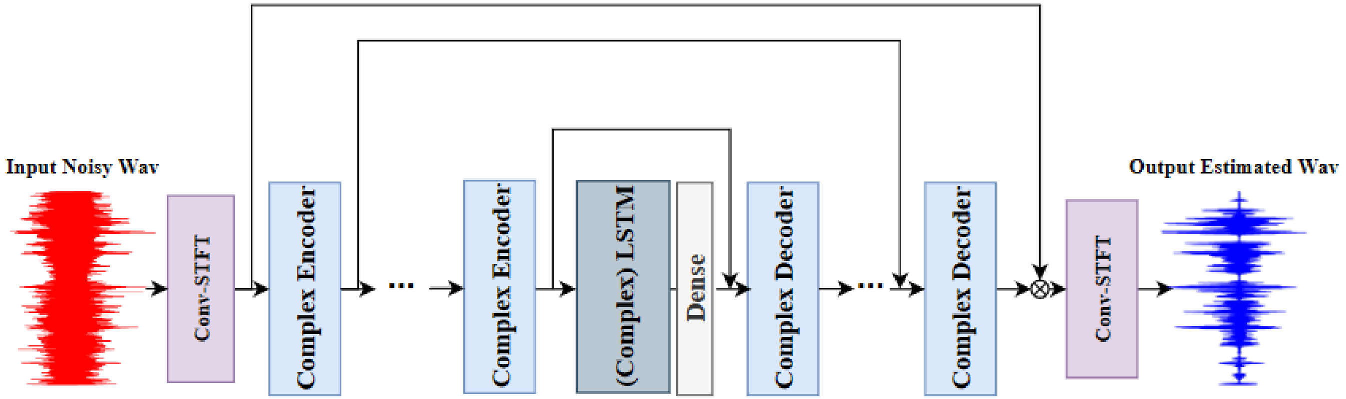
- **DCCRN (2020)**
  - Why not to use the STFT coefficient (with phases) directly?
  - Complex operations with complex data



(a) complex convolution



(b) complex encoder

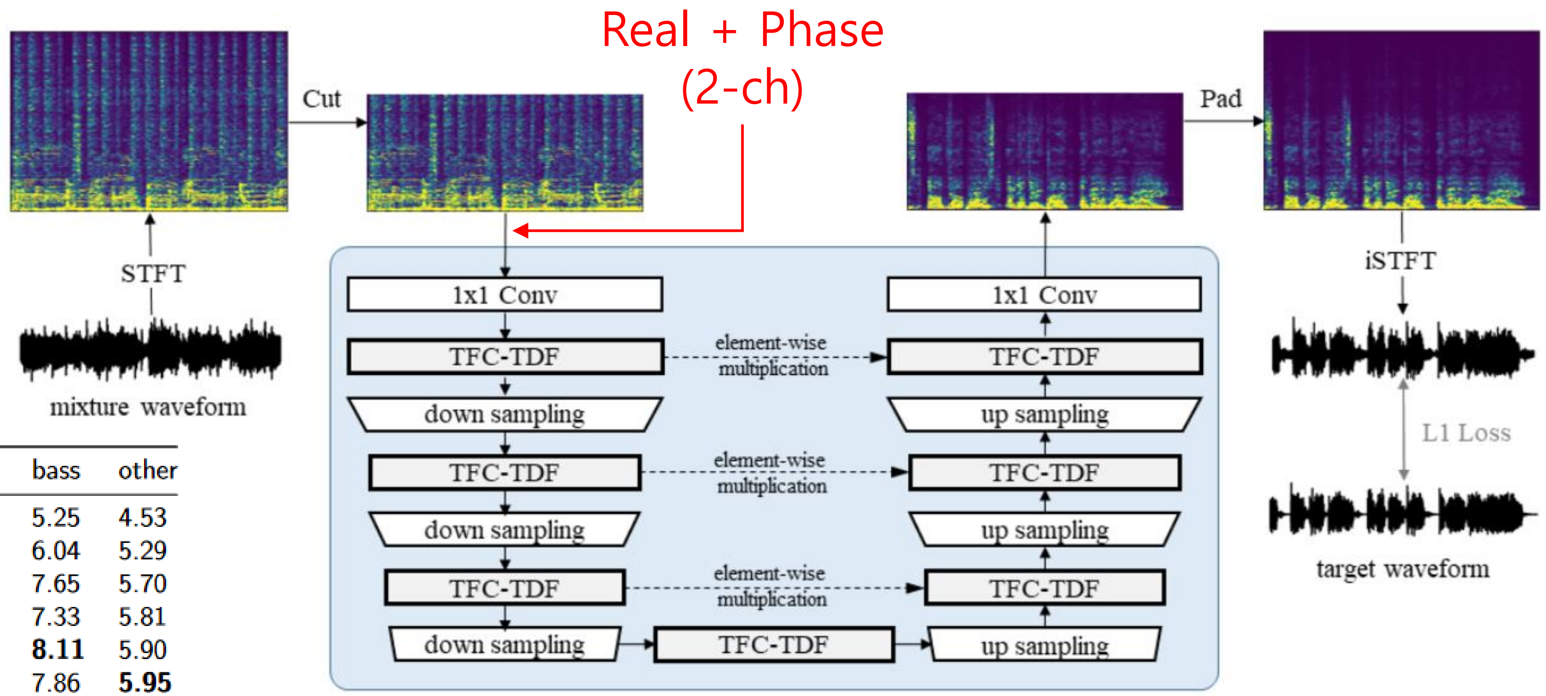


Better PESQ scores

Model	Para. (M)	look-ahead (ms)	no reverb	reverb	Ave.
Noisy	-	-	2.454	2.752	2.603
NSNet (Baseline) [34]	1.3	0	2.683	2.453	2.568
DCCRN-E [T1]	3.7	37.5	**3.266**	3.077	3.171
DCCRN-E-Aug [T2]	3.7	37.5	3.209	**3.219**	**3.214**
DCCRN-CL [T2]	3.7	37.5	3.262	3.101	3.181
DCUNET [T2]	3.6	37.5	3.223	2.796	3.001

MDX-Net (2021)

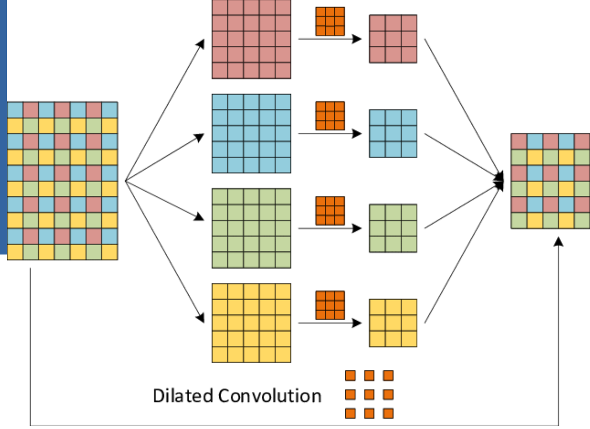
- **MDX-Net (2021)**
  - Complex convolution structure respecting Time and Frequency … (inherited from U-net ideas)
  - Separation model for each category + Mixer (즉, 각 소스별로 독립적으로 학습)



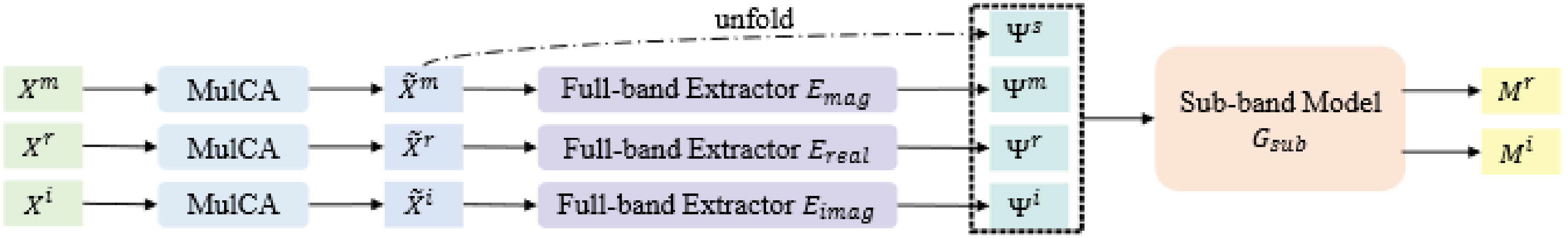
	vocals	drums	bass	other
D3Net (Takahashi & Mitsufuji, 2021)	7.24	7.01	5.25	4.53
ResUNetDecouple+ (Kong et al., 2021)	8.98	6.62	6.04	5.29
TFC-TDF-U-Net v2	8.81	6.52	7.65	5.70
v2 + Mixer	8.91	7.07	7.33	5.81
v2 + Demucs	8.80	7.14	**8.11**	5.90
KUIELab-MDX-Net	**9.00**	**7.33**	7.86	**5.95**

SNR of separation

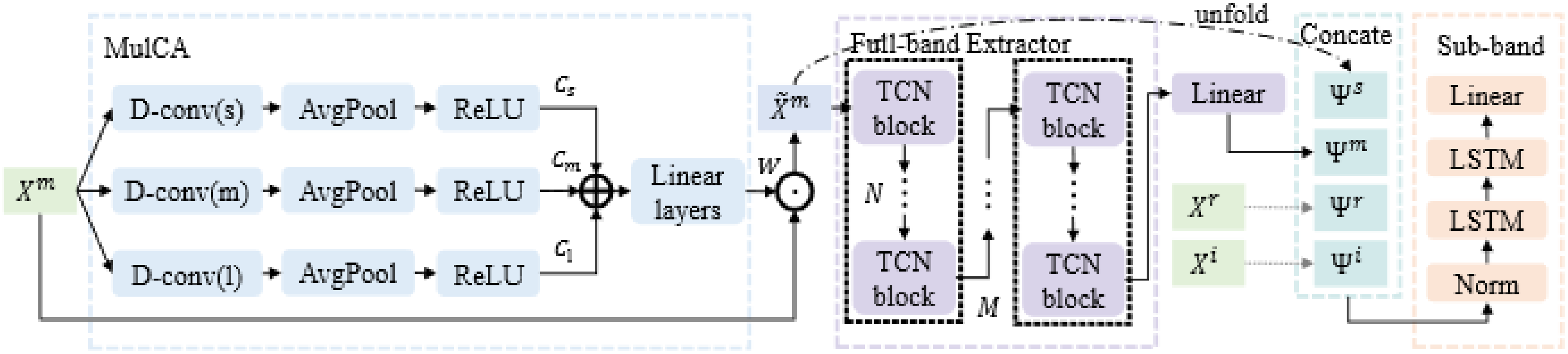
FullSubNet+(2022)



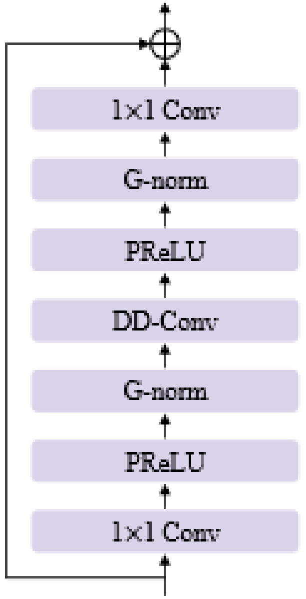
- **FullSubNet+(2022)**
  - **Idea**: use separately magnitude and 2-component phase, encode it via dilated convolutions then fully convolutional



(a) FullSubNet+ diagram



(b) System flowchart on one branch of the model



TCN-Block

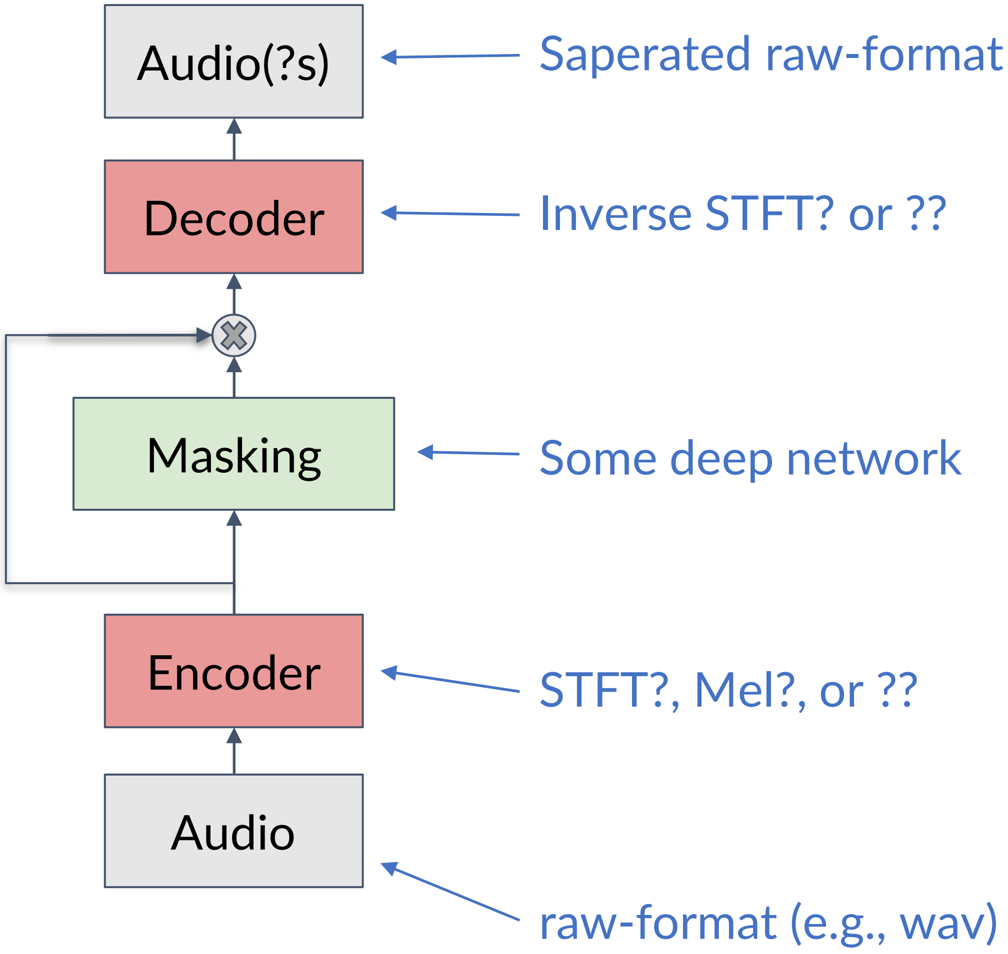
FullSubNet+(2022)

- **FullSubNet+(2022)**
  - **Idea**: use separately magnitude and 2-component phase, encode it via dilated convolutions then fully convolutional

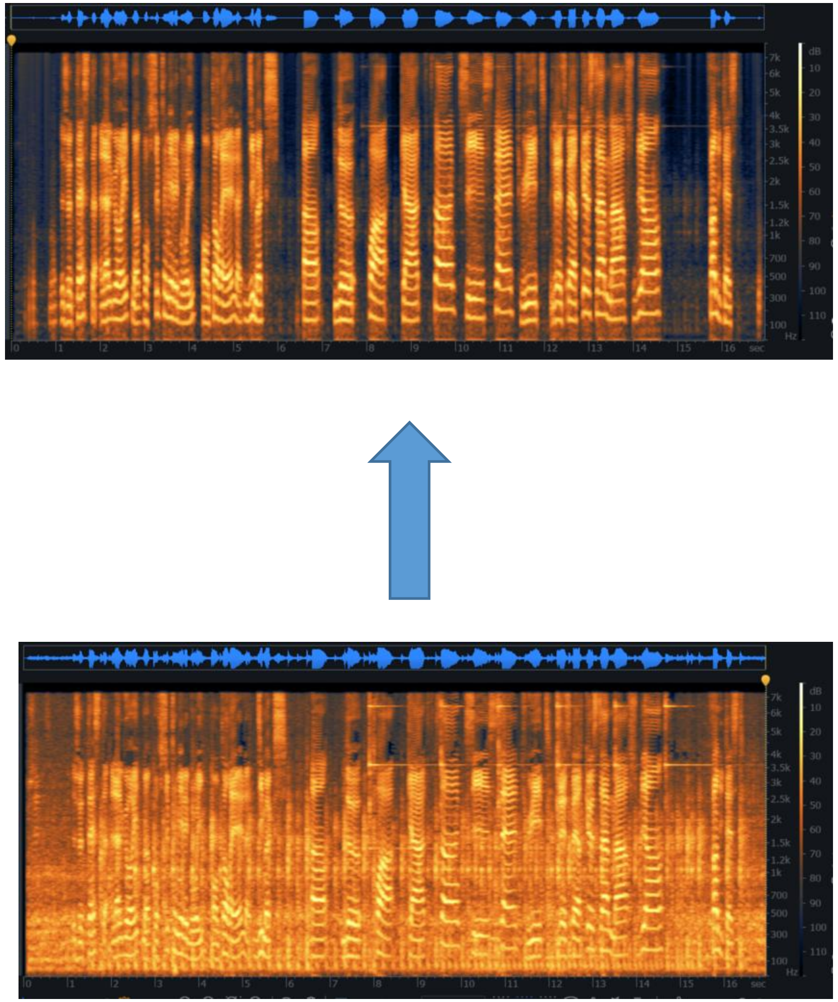
Table 1. The performance in terms of WB-PESQ [MOS], NB-PESQ [MOS], STOI [%], and SI-SDR [dB] on the DNS Challenge test dataset.

Model	Year	Look Ahead (ms)	With Reverb				Without Reverb			
			WB-PESQ	NB-PESQ	STOI	SI-SDR	WB-PESQ	NB-PESQ	STOI	SI-SDR
Noisy	-	-	1.822	2.753	86.62	9.033	1.582	2.454	91.52	9.07
DCCRN-E [22]	2020	37.5	-	3.077	-	-	-	3.266	-	-
PoCoNet [23]	2020	-	2.832	-	-	-	2.748	-	-	-
DCCRN+ [24]	2021	10	-	3.300	-	-	-	3.330	-	-
TRU-Net [25]	2021	0	2.740	3.350	91.29	14.87	2.860	3.360	96.32	17.55
CTS-Net [26]	2021	-	3.020	3.470	92.70	15.58	2.940	3.420	96.66	17.99
FullSubNet [12]	2021	32	3.063	3.581	92.93	16.09	2.813	3.403	96.17	17.44
FullSubNet+	2021	32	**3.218**	**3.666**	**93.84**	**16.81**	**2.982**	**3.504**	**96.69**	**18.34**

DEMUCS Denoiser (2020)

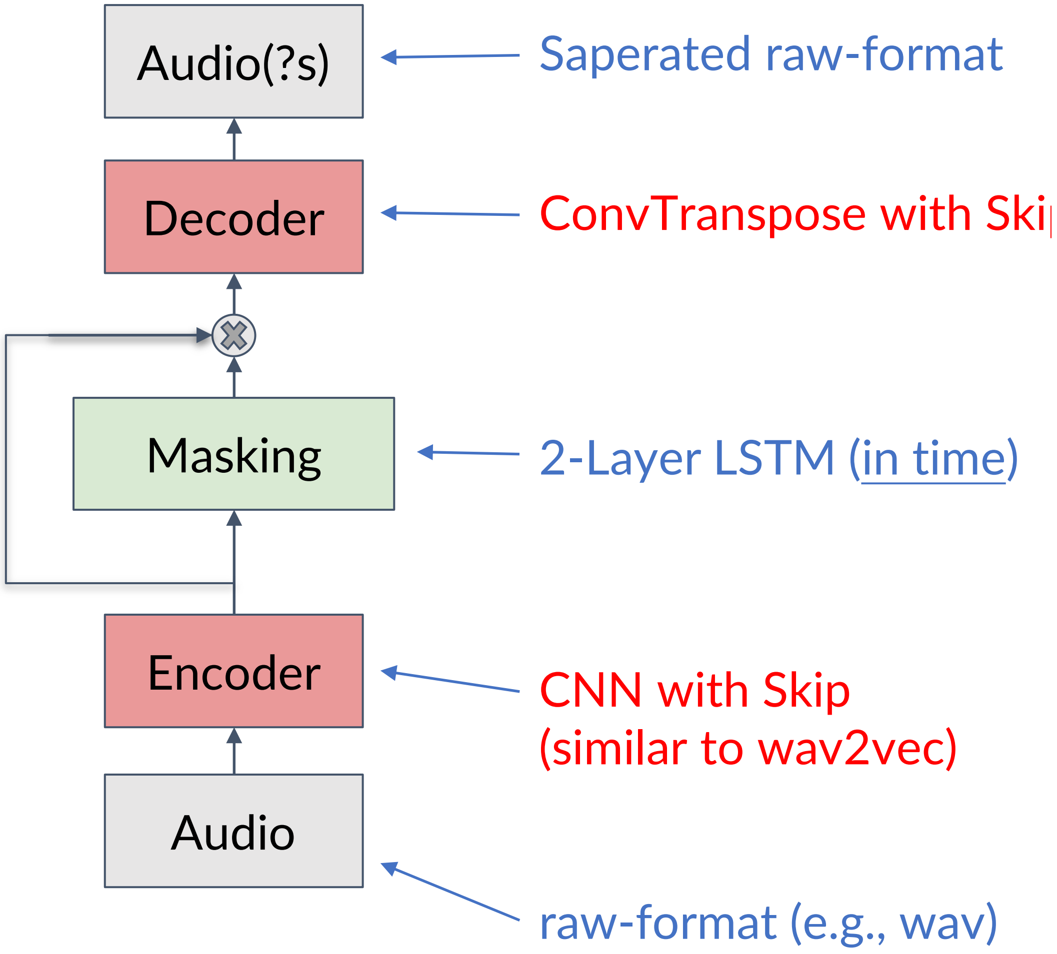


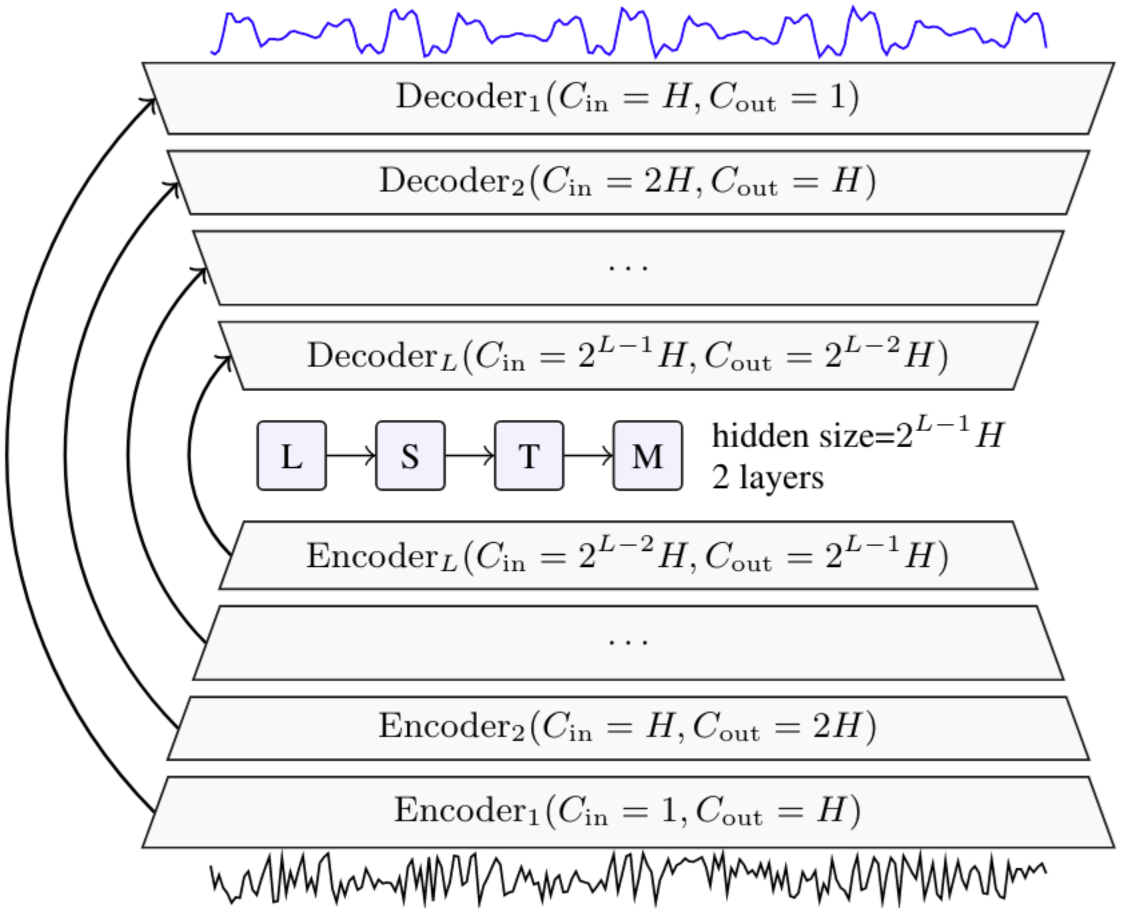
Denoised
Alex



Alex
&
noise

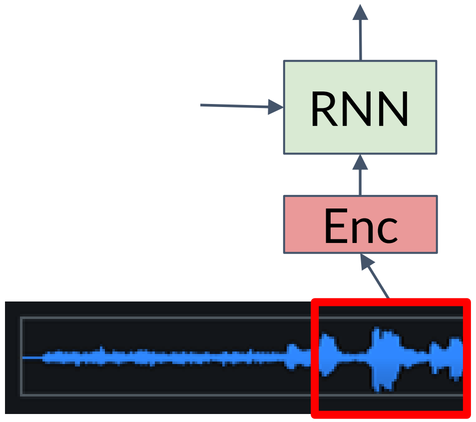
DEMUCS Denoiser (2020)



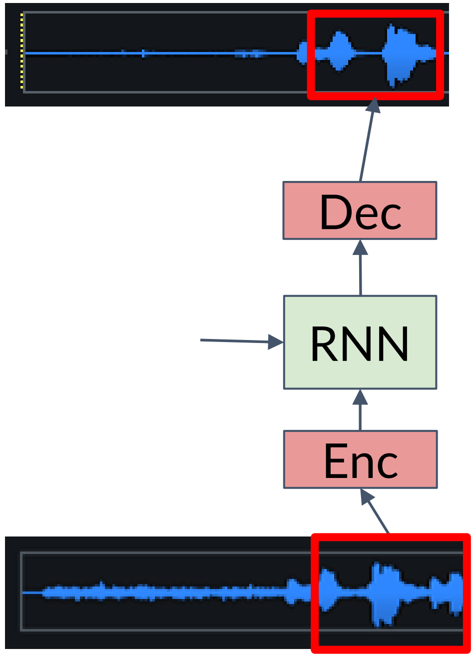


DEMUCS Denoiser (2020)

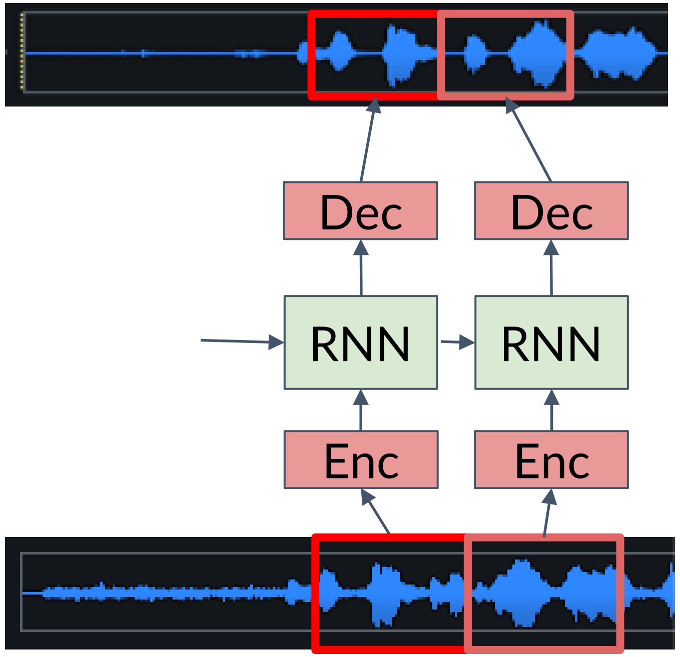
DEMUCS Denoiser (2020)



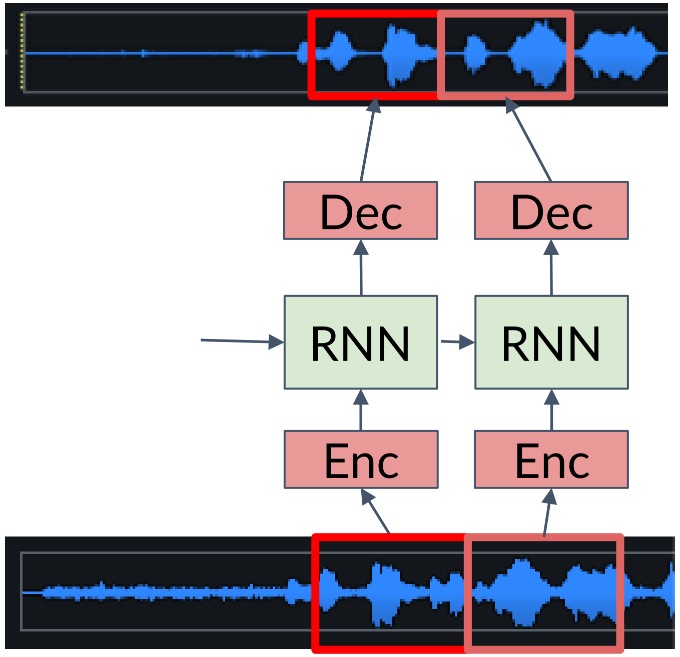
DEMUCS Denoiser (2020)



DEMUCS Denoiser (2020)

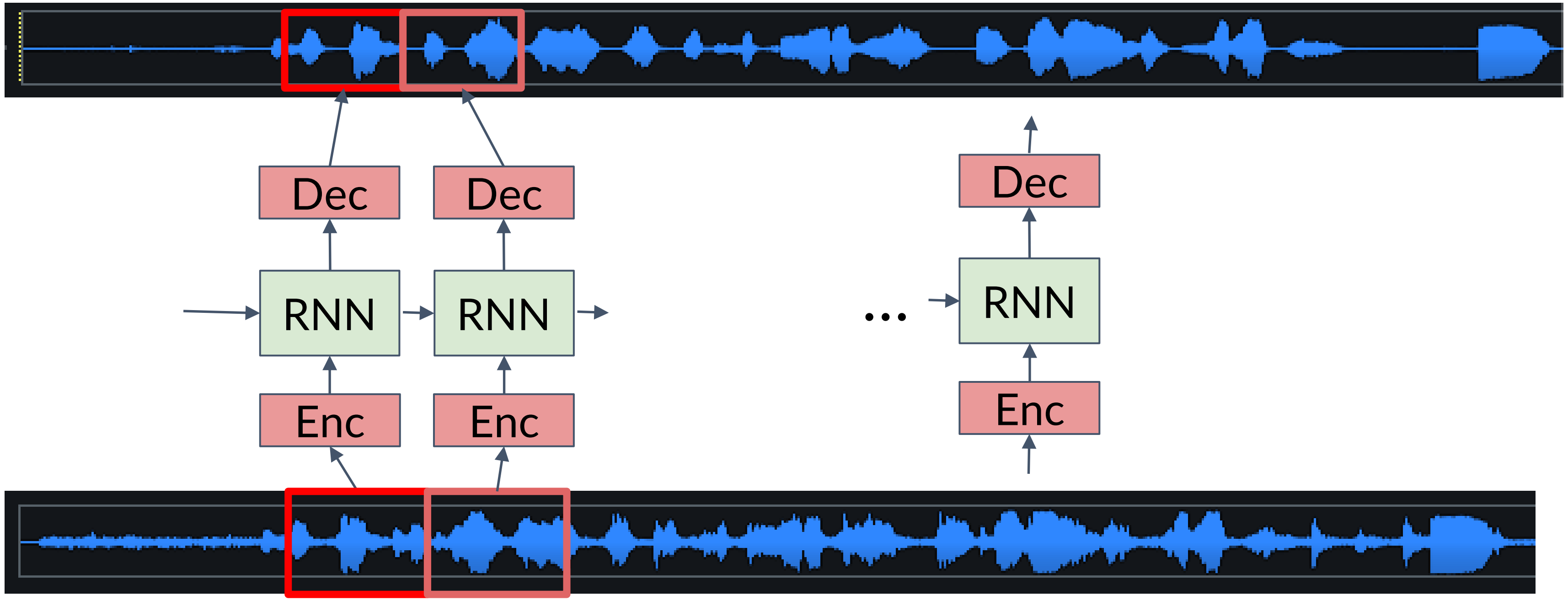


DEMUCS Denoiser (2020)



…

DEMUCS Denoiser (2020)

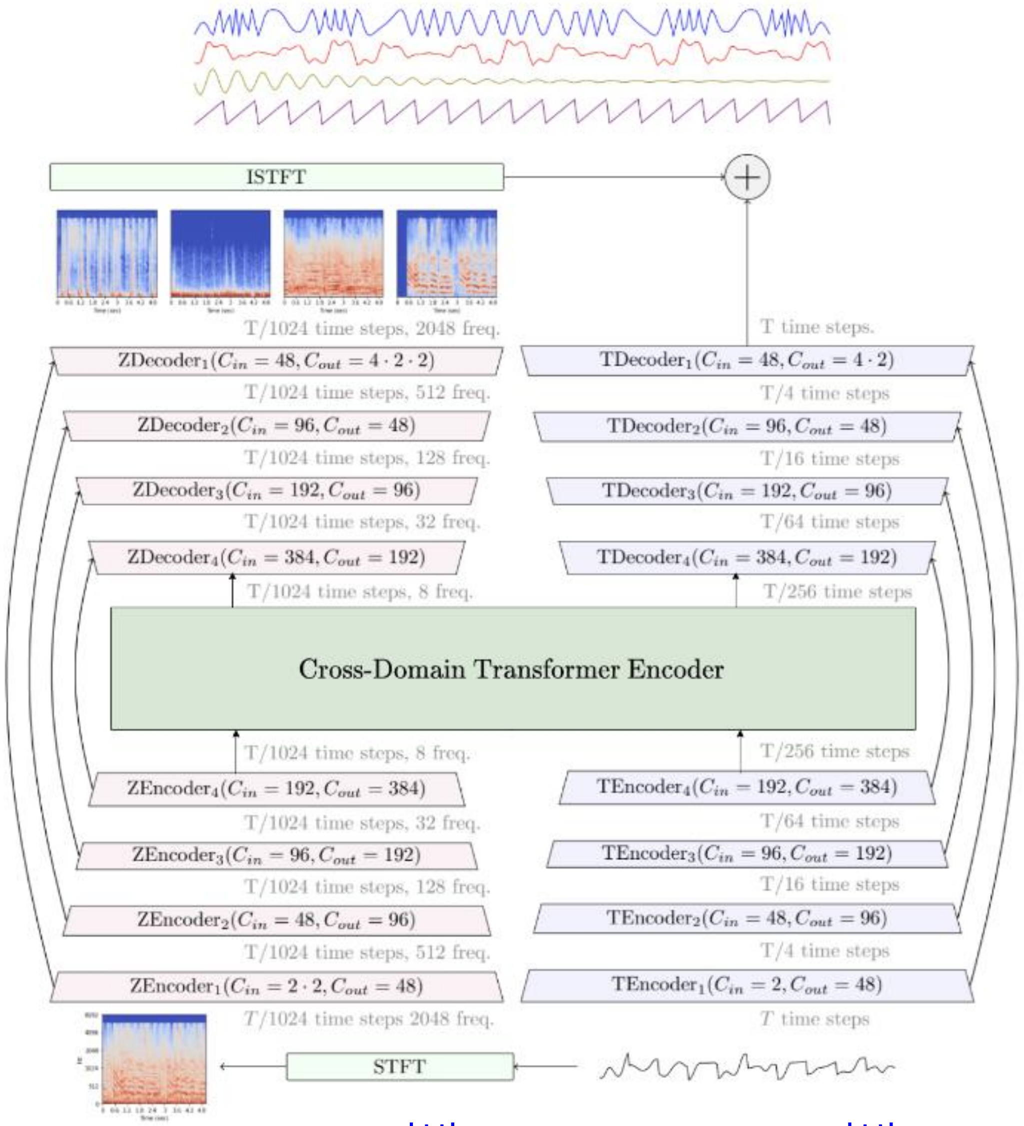


DEMUCS Denoiser (2020)

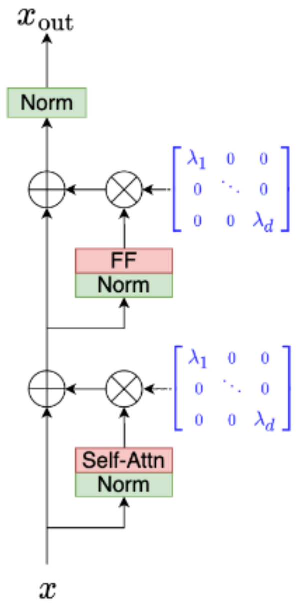
			Test SDR in dB				
Architecture	**Wav?**	**Extra?**	All	Drums	Bass	Other	Vocals
IRM oracle	✗	N/A	8.22	8.45	7.12	7.85	9.43
Wave-U-Net	✓	✗	3.23	4.22	3.21	2.25	3.25
Open-Unmix	✗	✗	5.33	5.73	5.23	4.02	6.32
Meta-Tasnet	✓	✗	5.52	5.91	5.58	4.19	6.40
Conv-Tasnet†	✓	✗	5.73 ±.10	6.02 ±.08	6.20 ±.15	4.27 ±.03	6.43 ±.16
DPRNN	✓	✗	5.82	6.15	5.88	4.32	6.92
D3Net	✗	✗	6.01	**7.01**	5.25	**4.53**	**7.24**
Demucs†	✓	✗	6.28 ±.03	6.86 ±.05	**7.01** ±.19	4.42 ±.06	6.84 ±.10
Spleeter	✗	∼ 25k*	5.91	6.71	5.51	4.55	6.86
TasNet	✓	∼ 2.5k	6.01	7.01	5.25	4.53	7.24
MMDenseLSTM	✗	804	6.04	6.81	5.40	4.80	7.16
Conv-Tasnet††	✓	150	6.32 ±.04	7.11 ±.13	7.00 ±.05	4.44±.03	6.74 ±.06
D3Net	✗	1.5k	6.68	7.36	6.20	**5.37**	**7.80**
Demucs†	✓	150	**6.79** ±.02	**7.58** ±.02	**7.60** ±.13	4.69 ±.04	7.29 ±.06

*: each track is only 30 seconds, †: from current work, ††: trained without pitch/tempo augmentation, as it deteriorates performance.

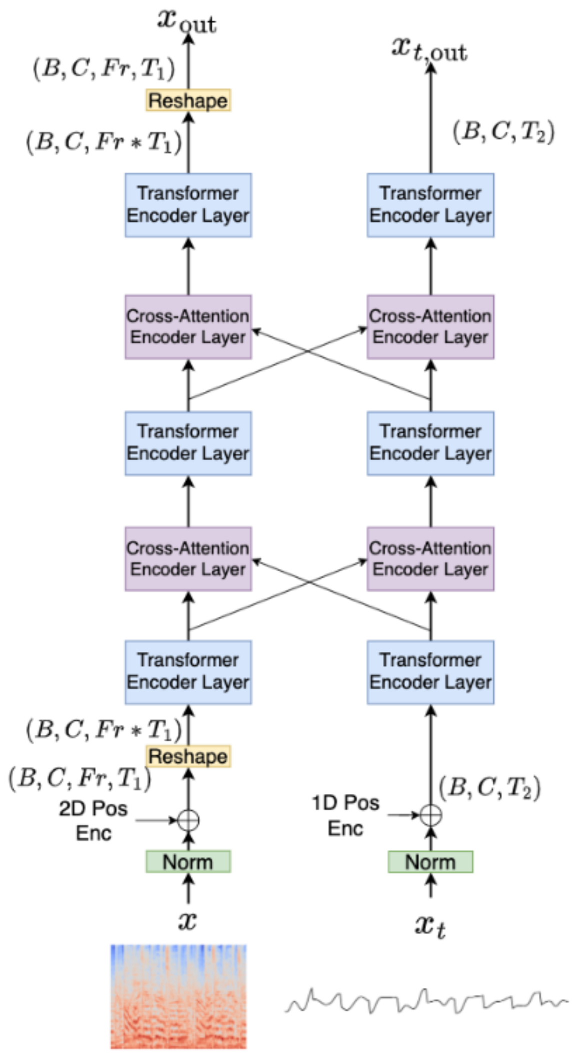
Hybrid Transformer DEMUCS (2022)



Spectrogram 기반 Waveform 기반



(a) Transformer Encoder Layer

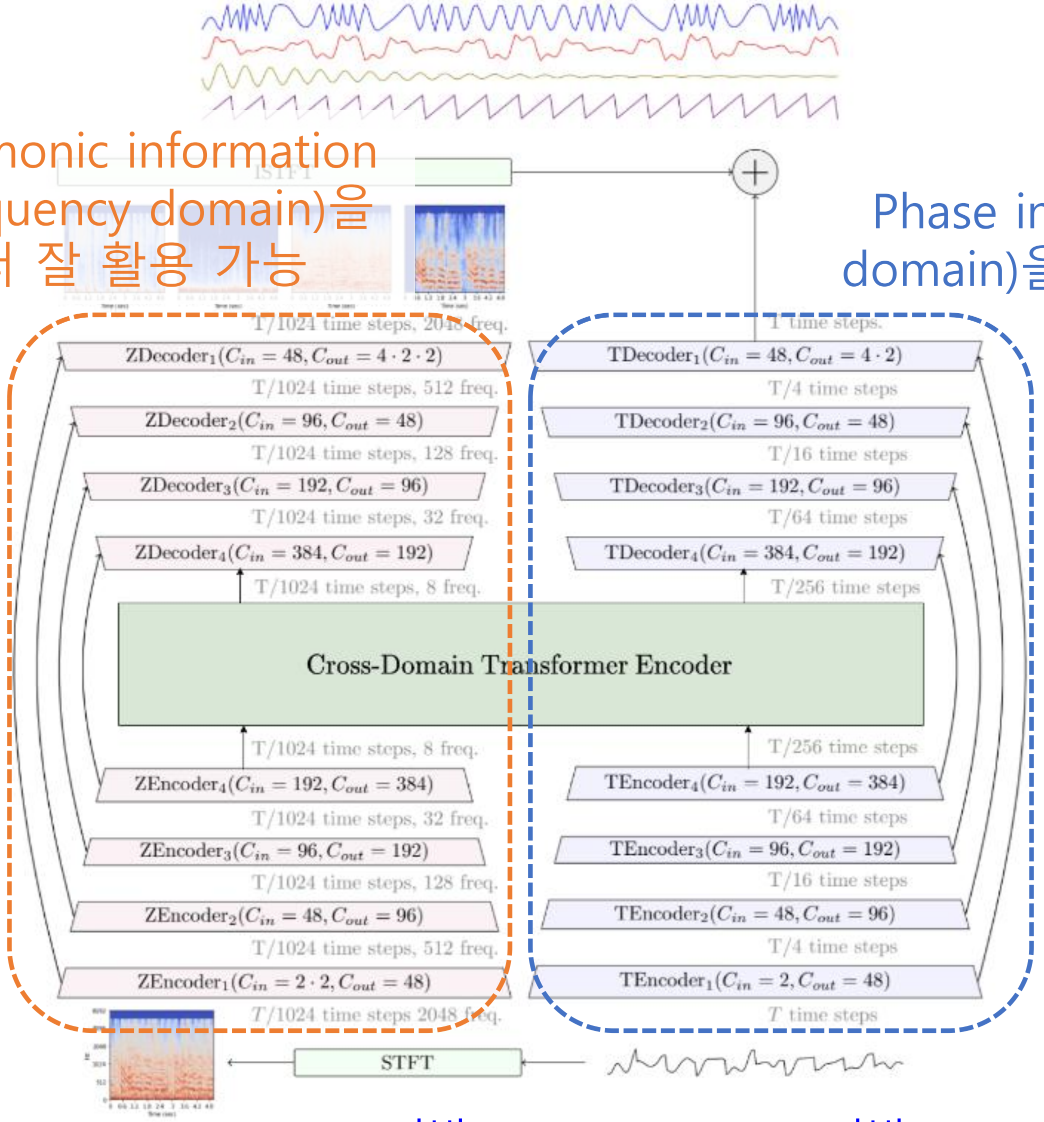


(b) Cross-domain Transformer Encoder of depth 5

Hybrid Transformer DEMUCS (2022)

Harmonic information (frequency domain)을 더 잘 활용 가능

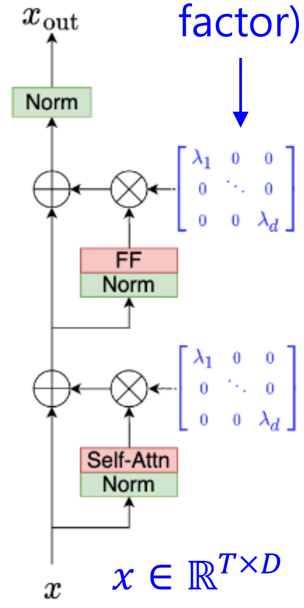
Phase information (time domain)을 더 잘 활용 가능



Spectrogram 기반

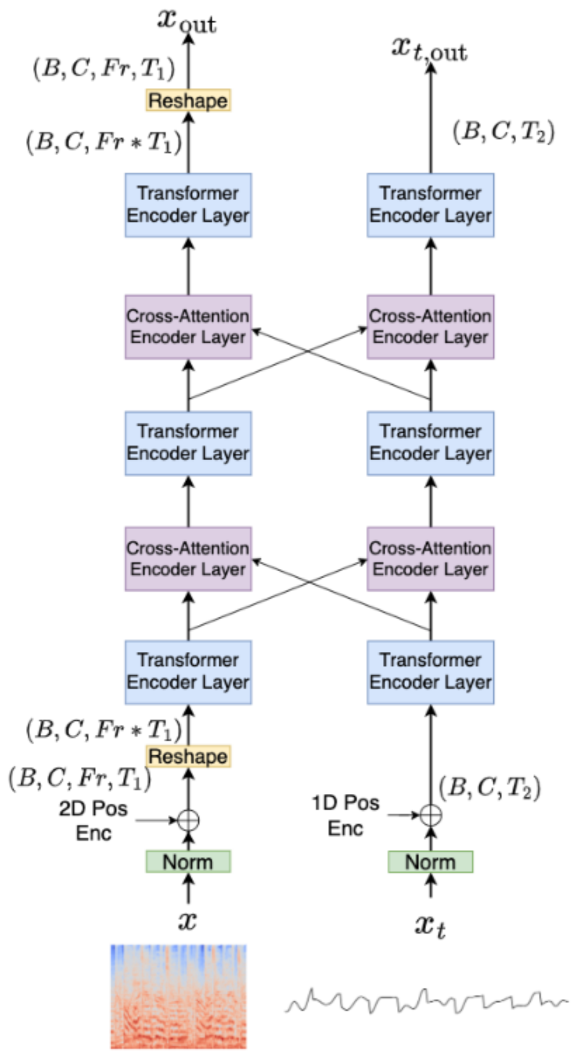
Waveform 기반

Diagonal matrix (adaptive scaling factor)



$x \in \mathbb{R}^{T \times D}$

(a) Transformer Encoder Layer



(b) Cross-domain Transformer Encoder of depth 5

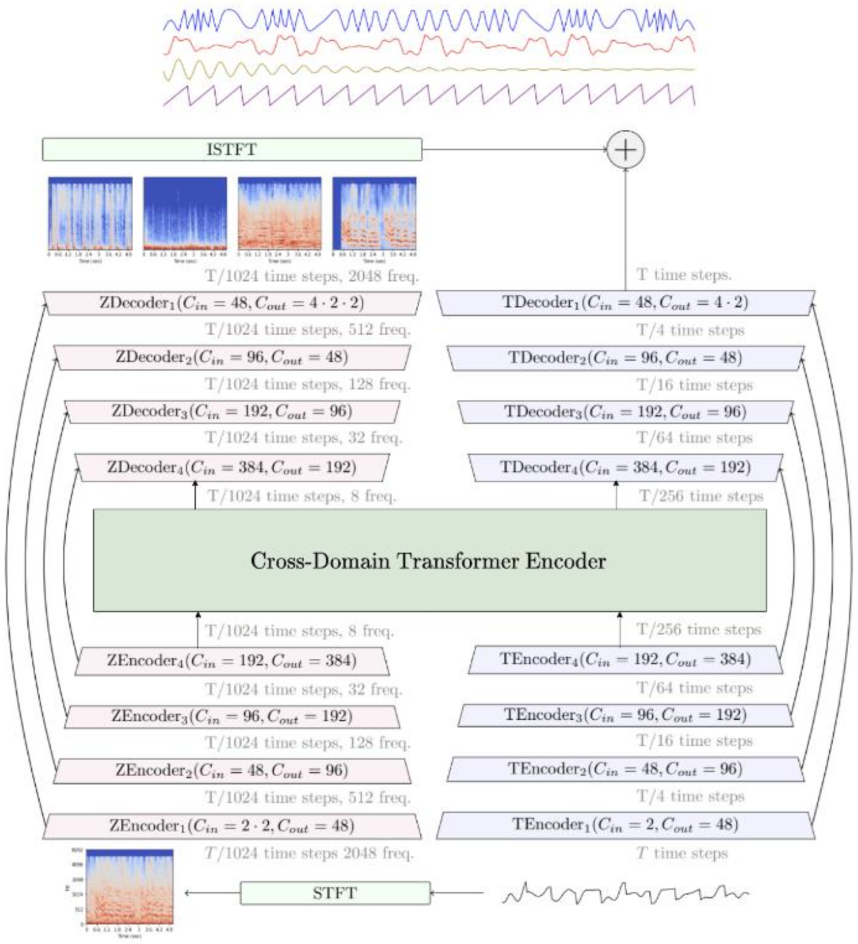
Hybrid Transformer DEMUCS (2022)

Table 3: Comparison on the MusDB (HQ for Hybrid Demucs) test set, using the original SDR metric. This includes methods that did not participate in the competition. "Mode" indicates if the waveform (W) or spectrogram (S) domain is used. Model with a "*" were evaluated on MusDB HQ.

Method	Mode	All	Drums	Bass	Other	Vocals
Hybrid Demucs*	S+W	**7.68**	**8.24**	**8.76**	5.59	8.13
Demucs v2	W	6.28	6.86	7.01	4.42	6.84
KUIELAB-MDX-Net*	S+W	7.47	7.20	7.83	**5.90**	**8.97**
D3Net	S	6.01	7.01	5.25	4.53	7.24
ResUNetDecouple+	S	6.73	6.62	6.04	5.29	**8.98**

BandSplit-RNN (2023)

- **More Conv, LSTM, ... ?**
  - DEMUCS has various versions: from **40M** to **86M** parameters
  - Release version is highly optimized
  - Reducing complexity through structed design: BSRNN (2023)

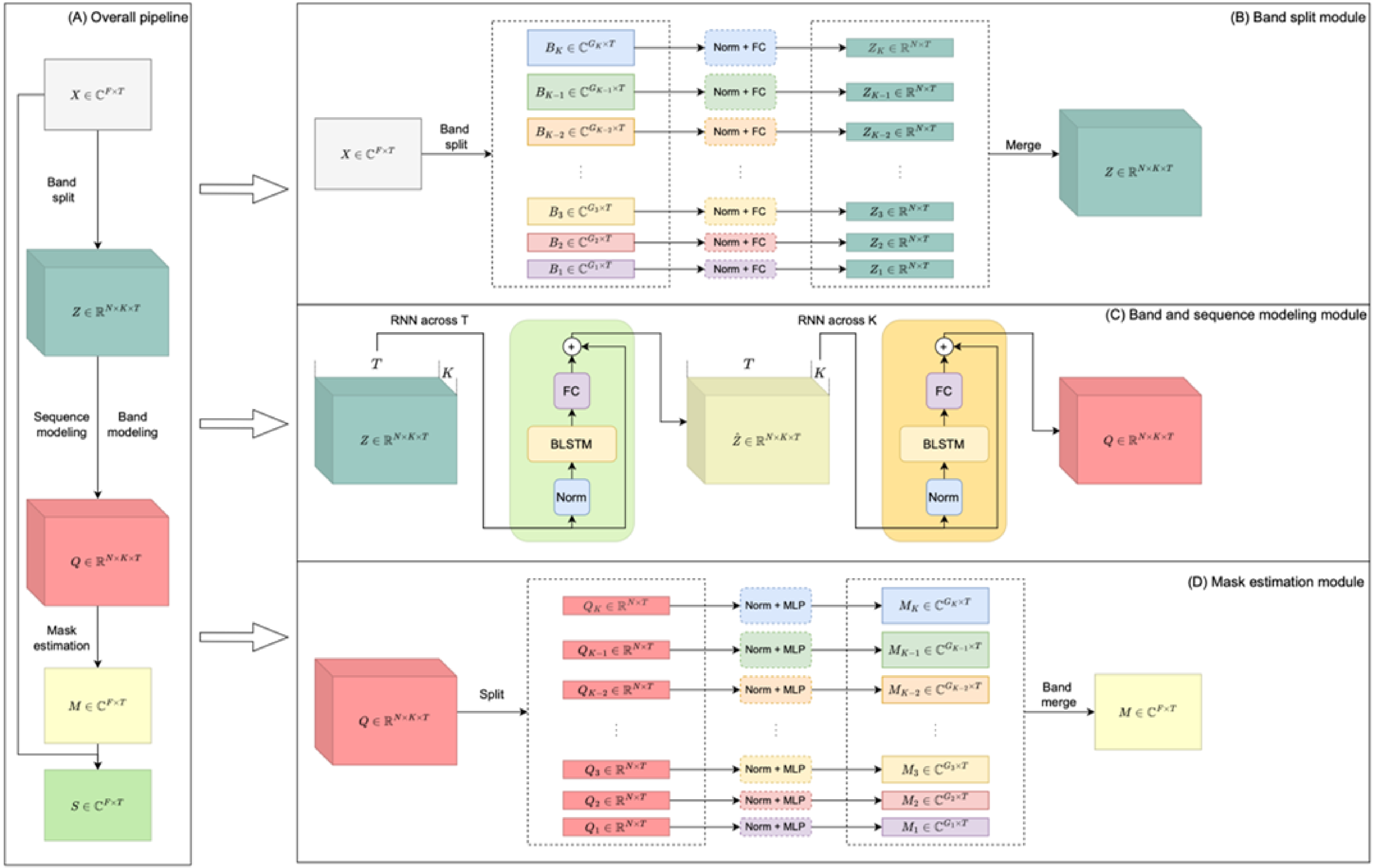


BandSplit-RNN (2023)

TF Spectrogram (complex) split into frequency bands

The bands are processed by Dual-Path RNN

Masks are reconstructed with band-wise MLP



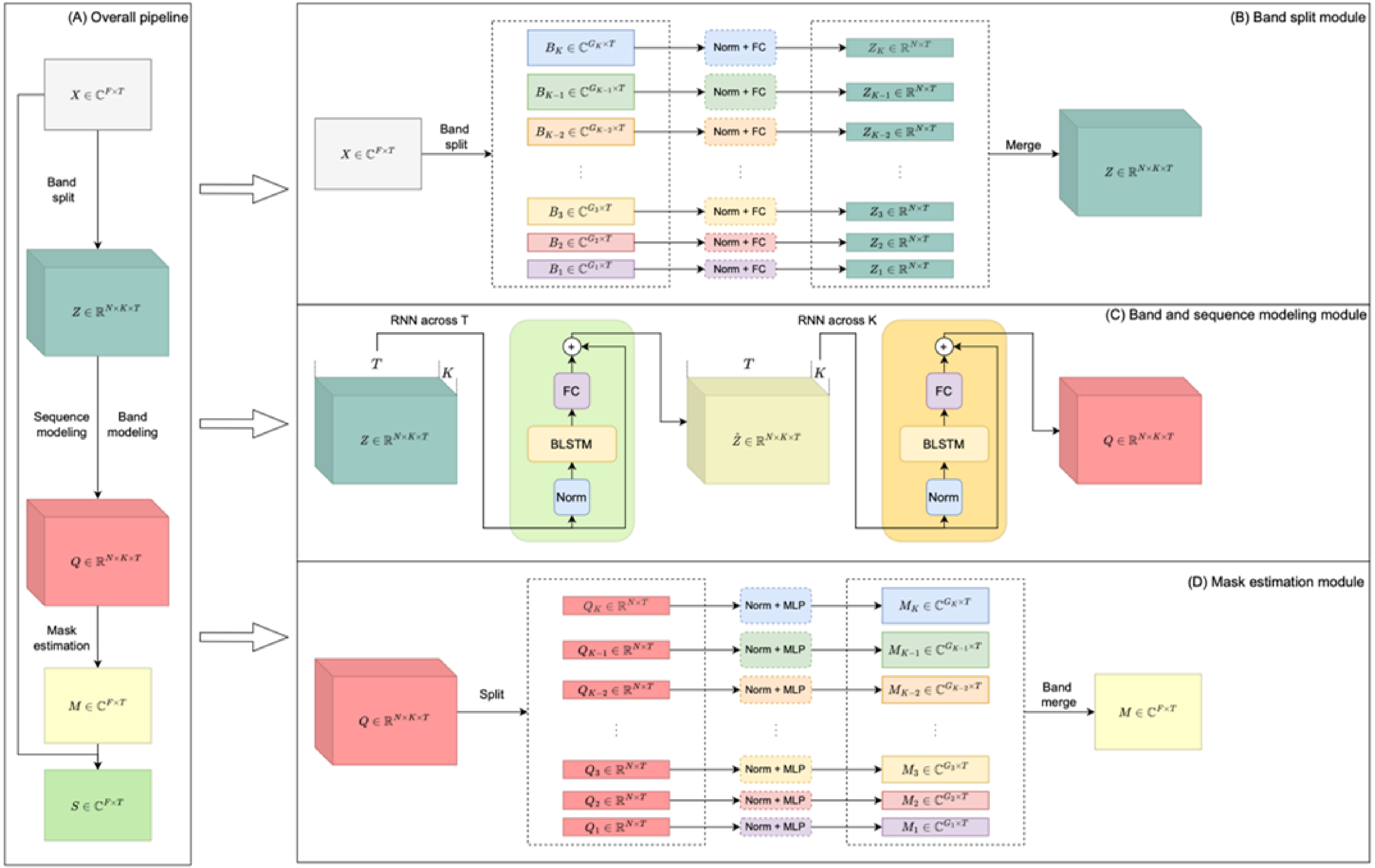
BandSplit-RNN (2023)

Semi-Supervised training + Supervised finetuning

TF Spectrogram (complex) split into frequency bands

The bands are processed by Dual-Path RNN

Masks are reconstructed with band-wise MLP



BandSplit-RNN（2023)

TABLE III. COMPARISON WITH EXISTING MODELS ON MUSDB18-HQ (HQ) AND MUSDB18 (NHQ) DATASET.

Model	Vocals				Bass				Drum				Other				All			
	uSDR		cSDR		uSDR		cSDR		uSDR		cSDR		uSDR		cSDR		uSDR		cSDR	
	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ	HQ	nHQ
ResUNetDecouple+ [25]	–	–	–	8.98	–	–	–	6.04	–	–	–	6.62	–	–	–	5.29	–	–	–	6.73
CWS-PResUNet [26]	–	–	8.92	–	–	–	5.93	–	–	–	6.38	–	–	–	5.84	–	–	–	6.77	–
KUIELab-MDX-Net [32]	–	–	8.97	9.00	–	–	7.83	7.86	–	–	7.20	7.33	–	–	5.90	5.95	–	–	7.47	7.54
Hybrid Demucs [31]	–	–	8.13	8.04	–	–	**8.76**	**8.67**	–	–	8.24	8.58	–	–	5.59	5.59	–	–	7.68	7.72
BSRNN	10.04	9.92	10.01	10.21	6.80	6.77	7.22	7.51	8.92	8.68	9.01	8.58	6.01	5.97	6.70	6.62	7.94	7.84	8.24	8.23
+ finetuning	**10.47**	**10.36**	**10.47**	**10.53**	7.20	7.17	8.16	8.30	**9.66**	**9.46**	**10.15**	**9.65**	**6.33**	**6.27**	**7.08**	**7.00**	**8.42**	**8.32**	**8.97**	**8.87**

~37M params per channel against 80M of DEMUCS