DL in Audio: Source Separation and more

NATIONAL RESEARCH UNIVERSITY

2024
Deep Learning in Audio Processing,
Invited Talks

Maksim Kaledin

Transforming the sound…

Source Separation literally means separate any source of particular interest…

Speech again but limited to

Separate signal (music, speech…) from noise

Denoising

Separate audio mix into separate tracks

Source Separation

Blind (any k, cocktail)

Guided (specific and known types of sources)

Target (specific source given by reference)

Guess when specific sound sources are active

Diarization

Transforming the sound…

Source Separation literally means separate any source of particular interest…

Speech again but limited to

Separate signal (music, speech…) from noise

Denoising

Separate audio mix into separate tracks

Source Separation

Blind (any k, cocktail)

Guided (specific and known types of sources)

Target (specific source given by reference)

Guess when specific sound sources are active

Diarization

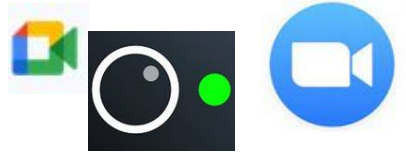
Applications

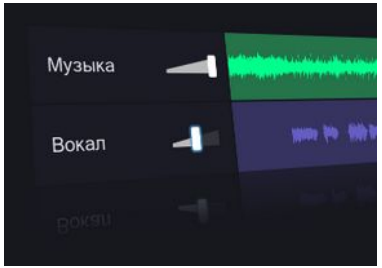
Source Separation literally means separate any source of particular interest…

Conference call systems (aka Zoom, Yandex Telemost, Google Meet etc.)

Audio Production (denoising the instrument or talking recordings; Izotope and more...)

Reconstructing the played sound for musician and composer routines (vocalremover, mvsep, lalai, moisesai…..)





Applications

Podcast Recording

Ref.: Сычёв подкаст

Applications

Podcast Recording

Just one stereo mic?..
Instead of studio?..

Ref.: Сычёв подкаст

Applications

Online Meetings

Ref.: Café Nuage

Denoising Problem

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

Early (and still popular) solutions: different types of spectral profiling

Main idea: the noise is different, we'll just cut it from the spectrum

Sounds simple…

[more history on Habr]

Ref.: Izotope RX 8

Denoising Problem

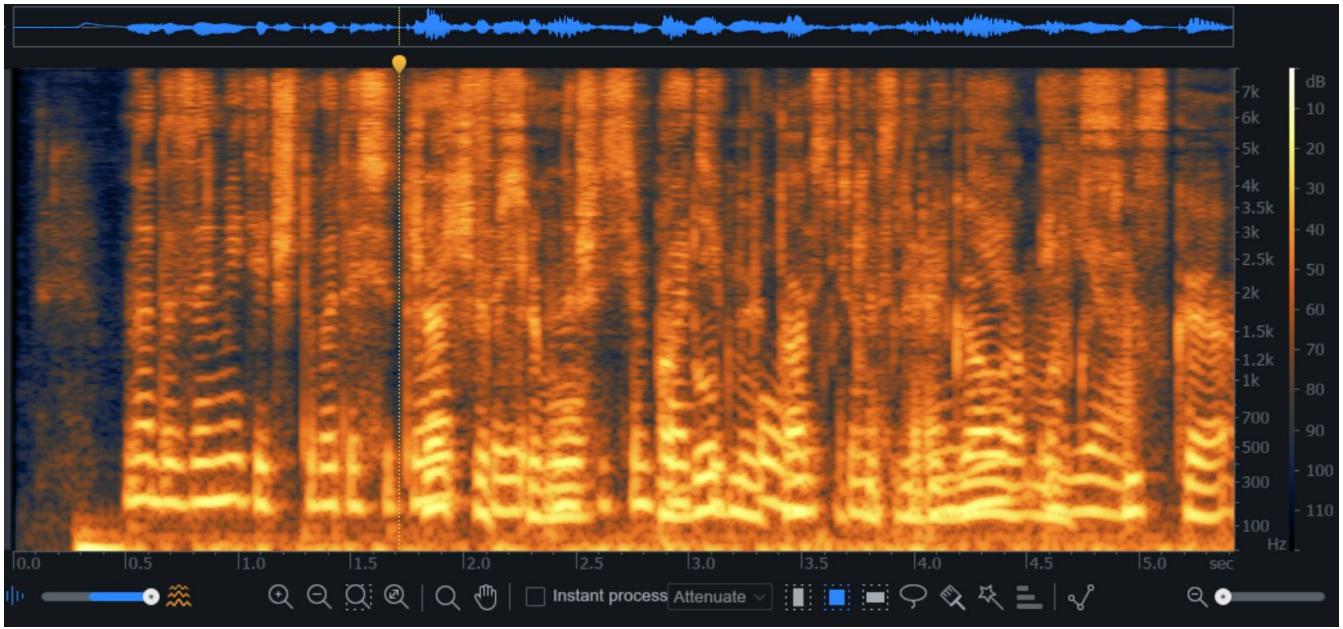
Goal: guess what is noise and remove the noise from audio

Early (and still popular) solutions: different types of spectral profiling

Main idea: the noise is different, we'll just cut it from the spectrum

The noise may not be well-separable !

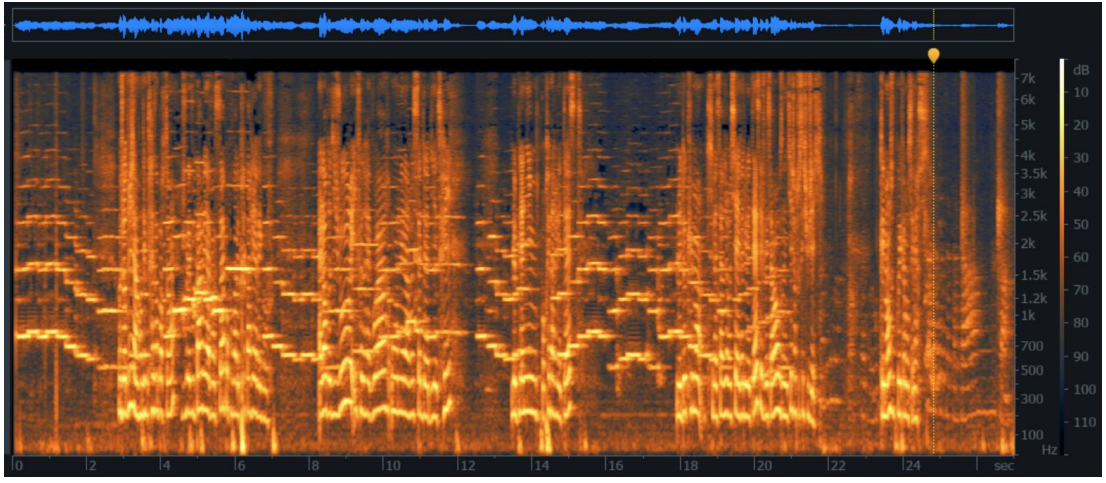
Two people speaking



Denoising Problem

Sarah & Flute

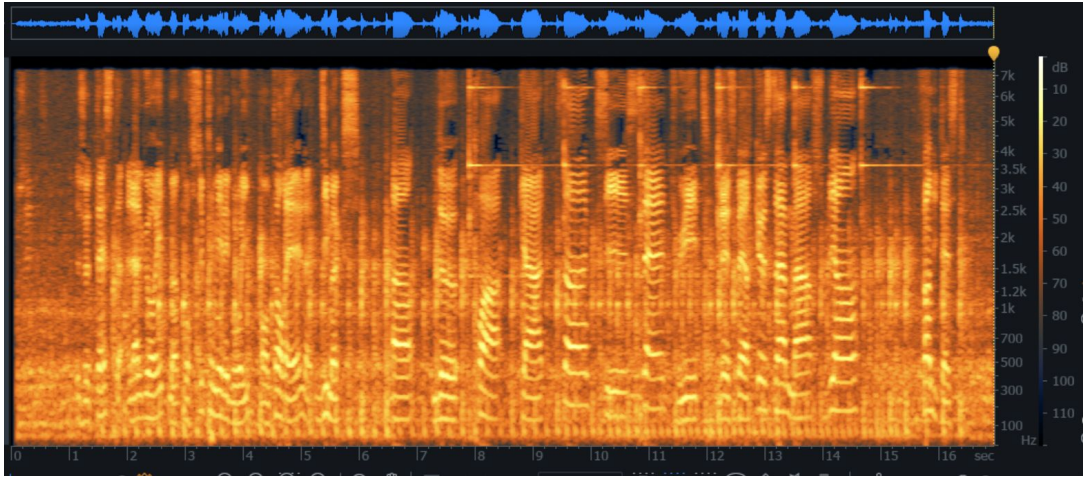
Deep Learning approach: see what can be done on the spectrogram



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

Alex & noise

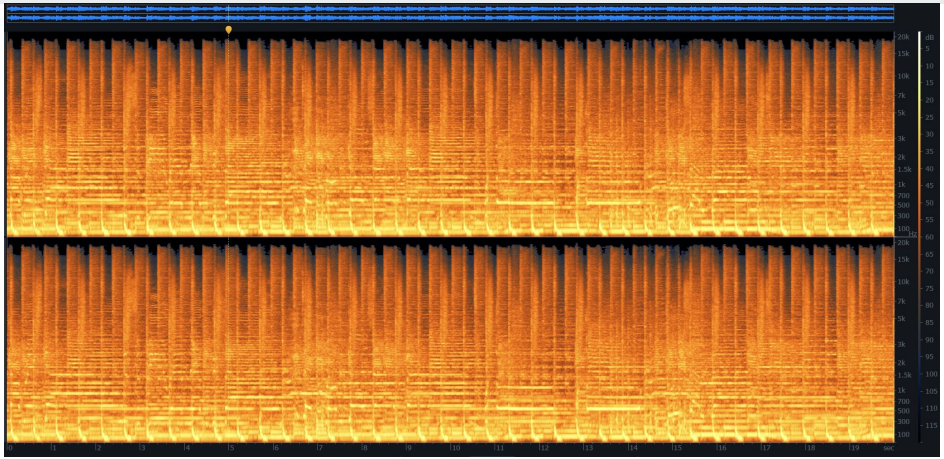
In a nontrivial way…



Audio Ref: DEMUCS

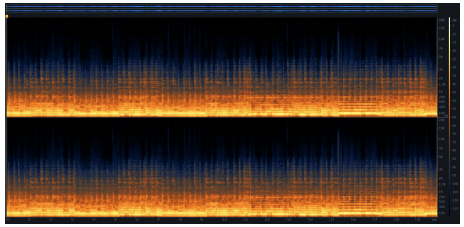
Separation Problem

Song

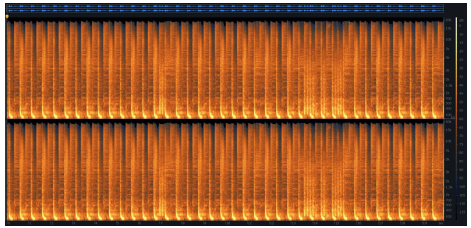


Deep Learning approach: very similar, just multiple audio in the end

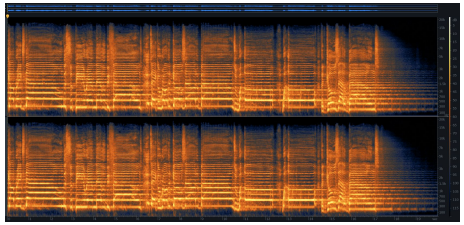
Main idea: we'll still just carve it from the "spectrum"



Bass



Drums



Vocals

Other

In a nontrivial way…

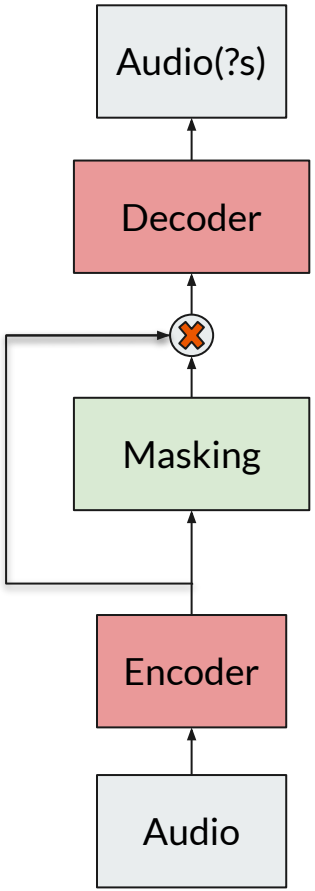
Audio Ref: DEMUCS, **The Easton Ellises - Falcon 69 (CC BY-NC SA)**

Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

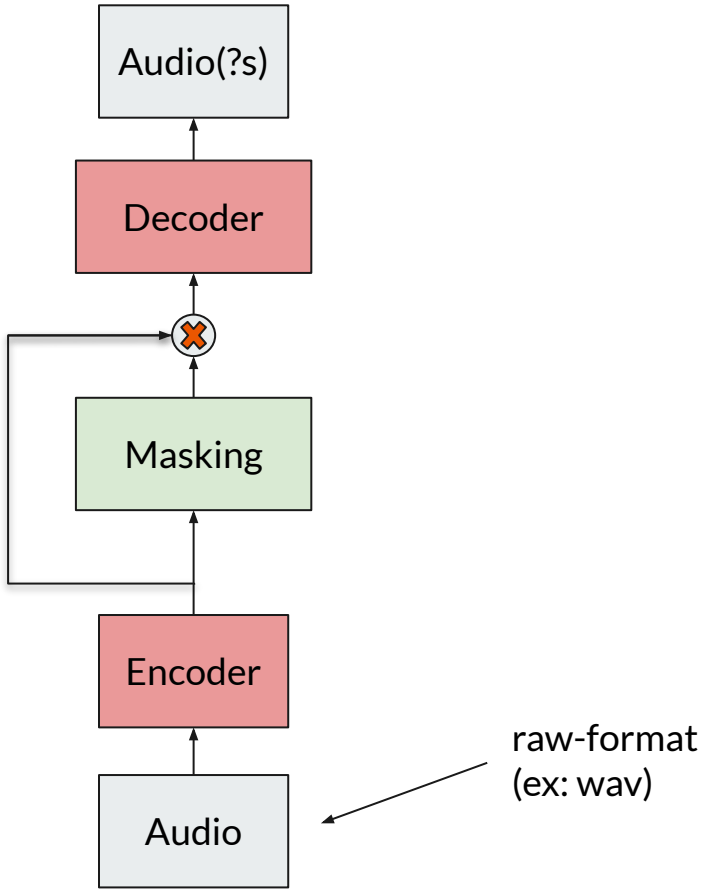


Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

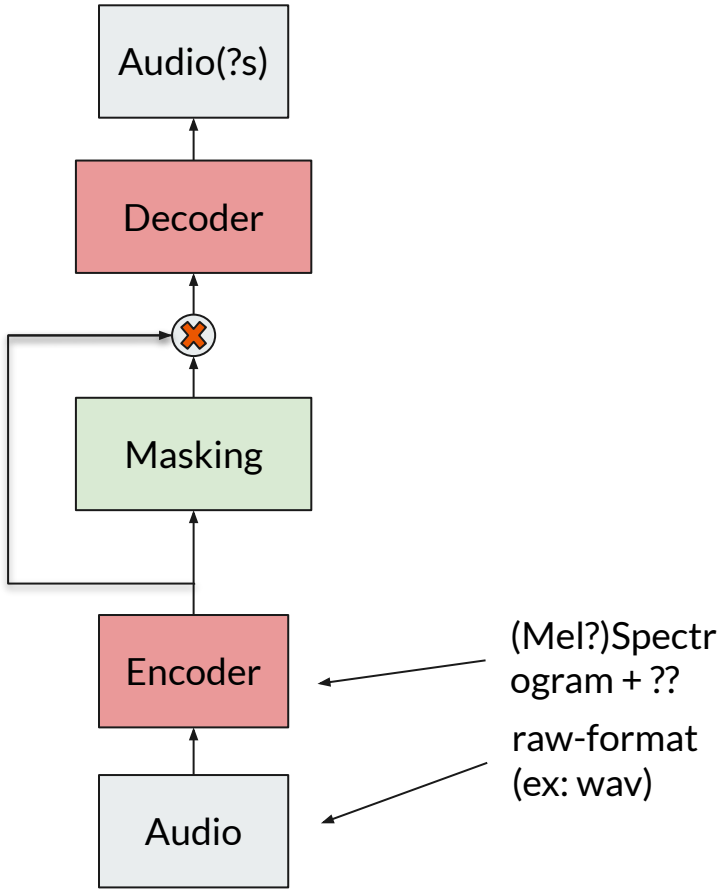


Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

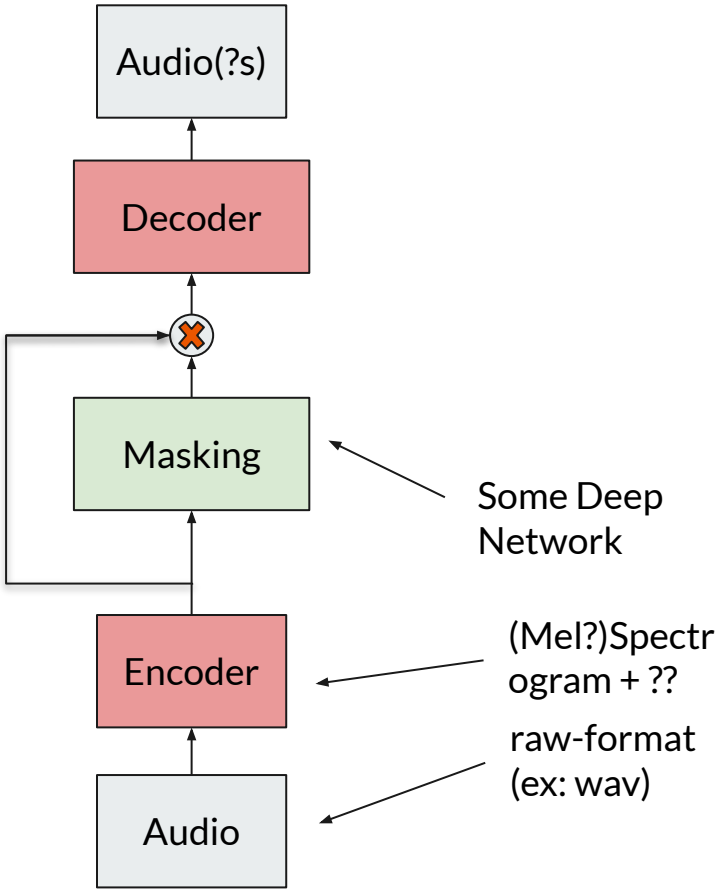


Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

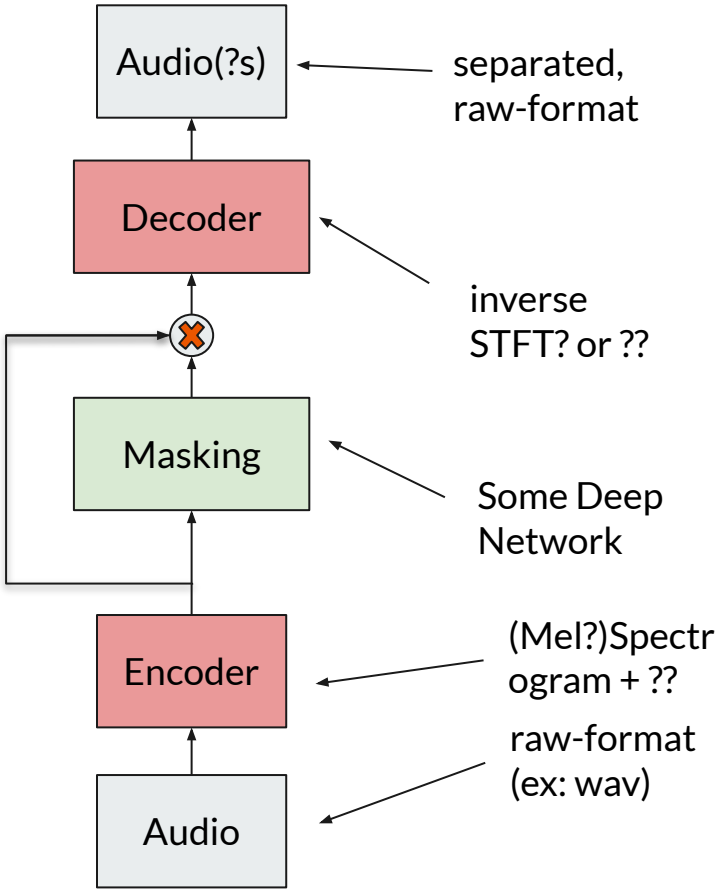


Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

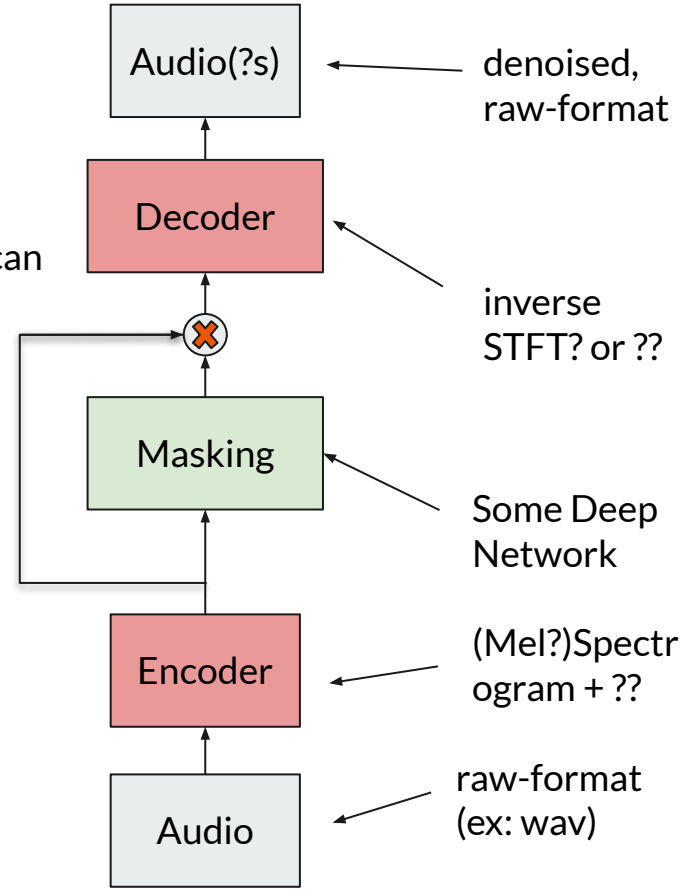


Denoising and Separation Pipeline

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…



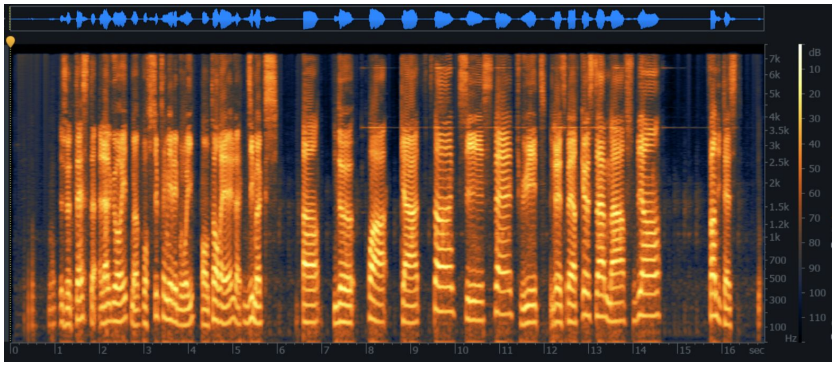
denoised, raw-format

inverse STFT? or ??

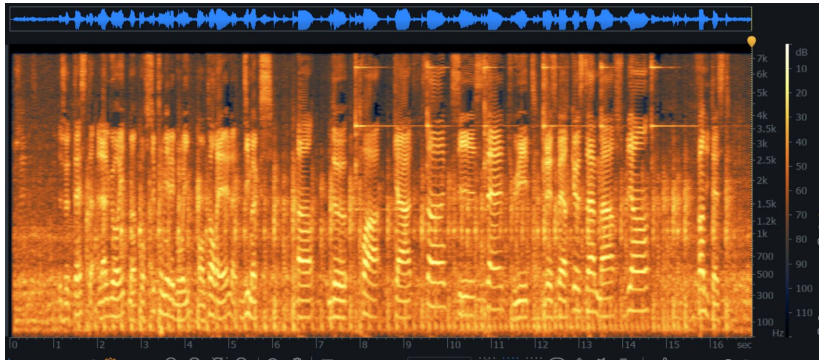
Some Deep Network

(Mel?)Spectr ogram + ??

raw-format (ex: wav)



Denoised Alex



Noisy Alex

Denoising and Separation Metrics

1. Signal-to-Noise Ratio (SNR, in the field same as SDR), in dB: $SDR(y, \hat{y}) = 10 \log_{10} \frac{\|y\|}{\|\hat{y} - y\|}$
2. SI-SNR (or SI-SDR), scale-invariant SNR, in dB
3. PESQ (**Perceptual Evaluation of Speech Quality**), expert-like, -0.5(bad) to 4.5(great)
4. STOI (**Short-Time Objective Intelligibility (STOI)**), expert-like, 0(bad) to 1(great)
5. MOS (Different mean opinion scores with variations)

Denoising and Separation Datasets

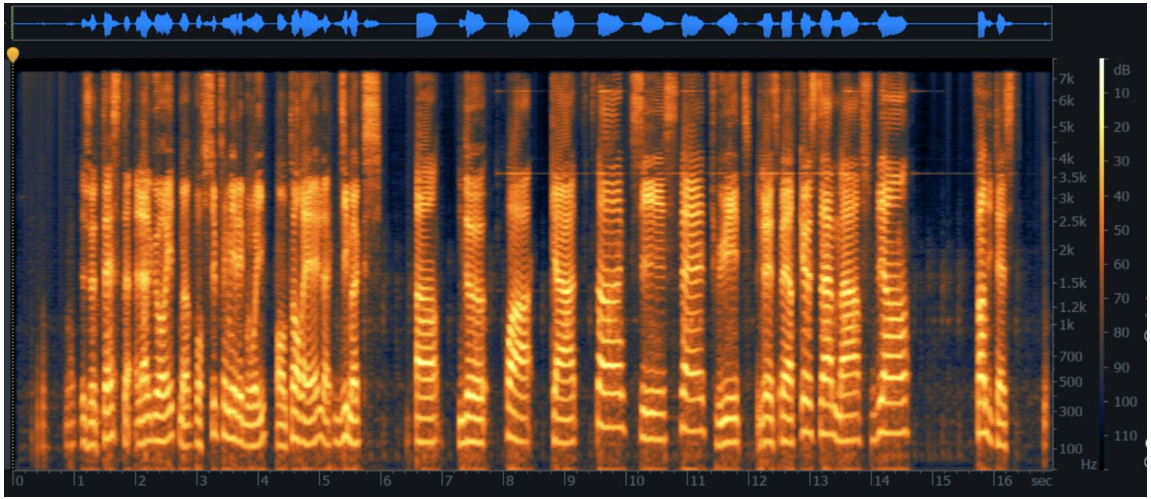
1. Speech: LibriSpeech(+noised), MS-SNSD, WSJ, LJSpeech ...
2. Music: MUSDB18 and MUSDB18HQ , DSD100 , Slakh (music tracks with stem tracks available) (2018-2020)
3. Noises: WHAM! and WHAMR! (2018), ARCA23K (2021)
4. Musical instruments with noise: FSDnoisy18k (2019)
5. Various stuff: VGGSound (2020)
6. Specific: DNS Challenges (2018-now)

.....

Spectrogram and Phase

Spectrogram information

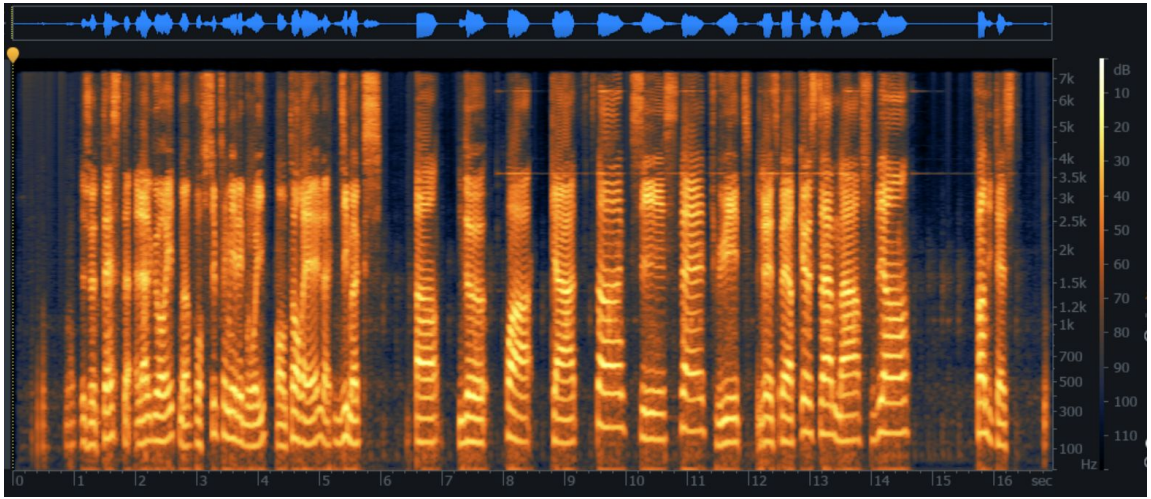
Spectrogram (just Amplitude info)



Spectrogram information

Idea: use spectrogram as part of the encoder and inverse STFT as a decoder from masked spectrogram.

Spectrogram (just Amplitude info)

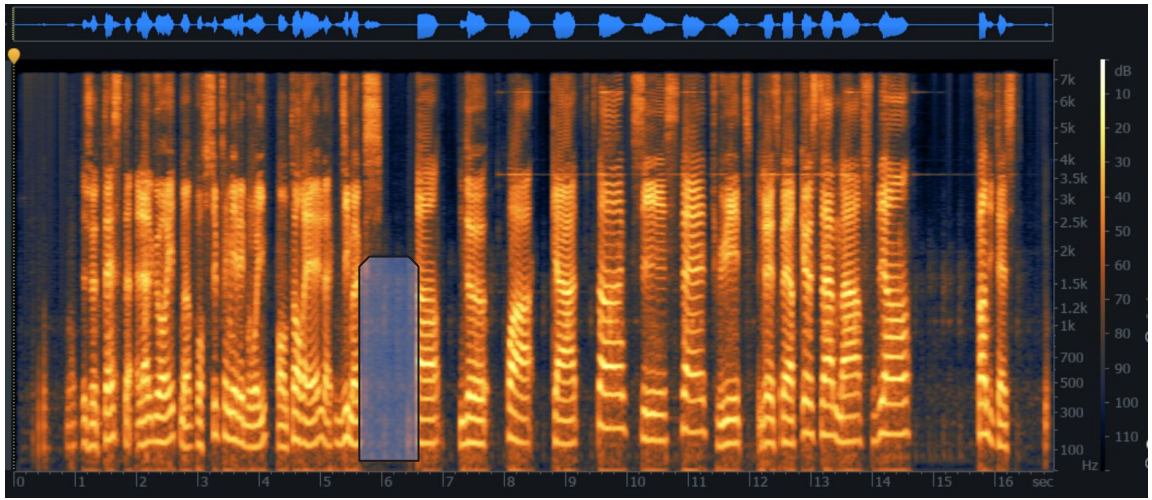


Spectrogram information

Idea: use spectrogram as part of the encoder and inverse STFT as a decoder from masked spectrogram.

Questions: loss?..
Spectrogram encoding?

Spectrogram (just Amplitude info)



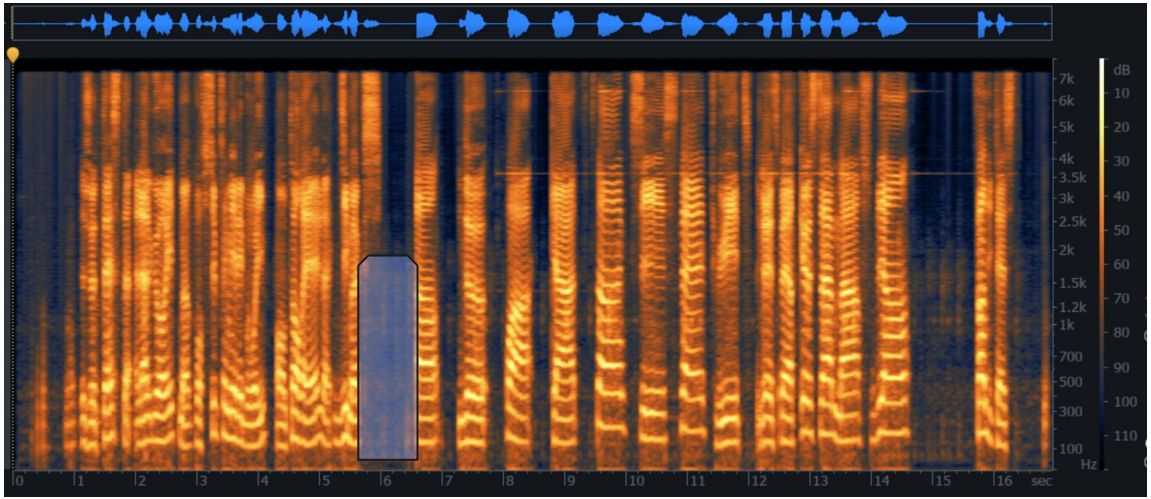
cut!

Spectrogram information

Idea: use spectrogram as part of the encoder and inverse STFT as a decoder from masked spectrogram.

Questions: loss?..
Spectrogram encoding?
Phase info?

Spectrogram (just Amplitude info)

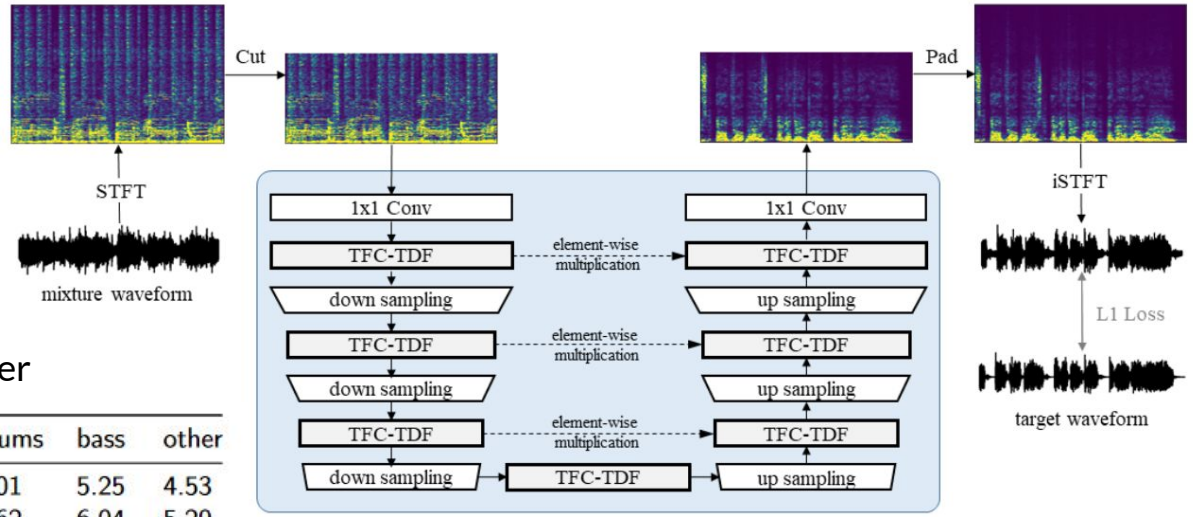


cut!

MDX-NET (2021)

Presented on MDX Workshop (2021)

Complex convolution structure respecting Time and Frequency… (inherited from U-Net ideas)



Separate model for each category + Mixer

	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**

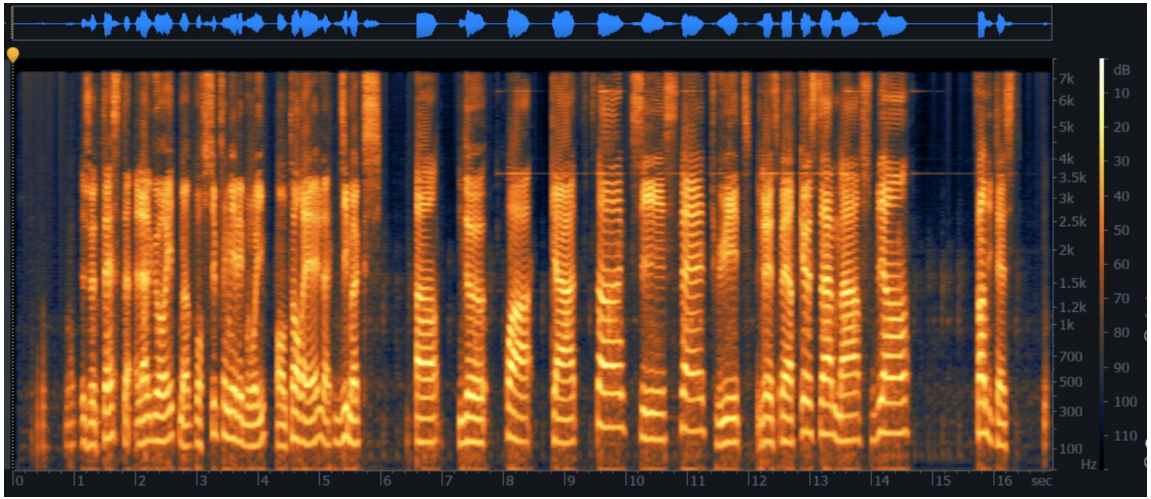
Ref: KUIELab-MDX-Net

SDR of separation

Spectrogram information

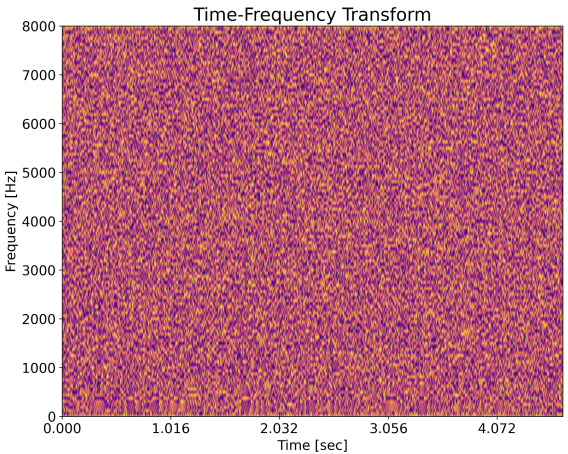
Can we use phase?..

Spectrogram (just Amplitude info)



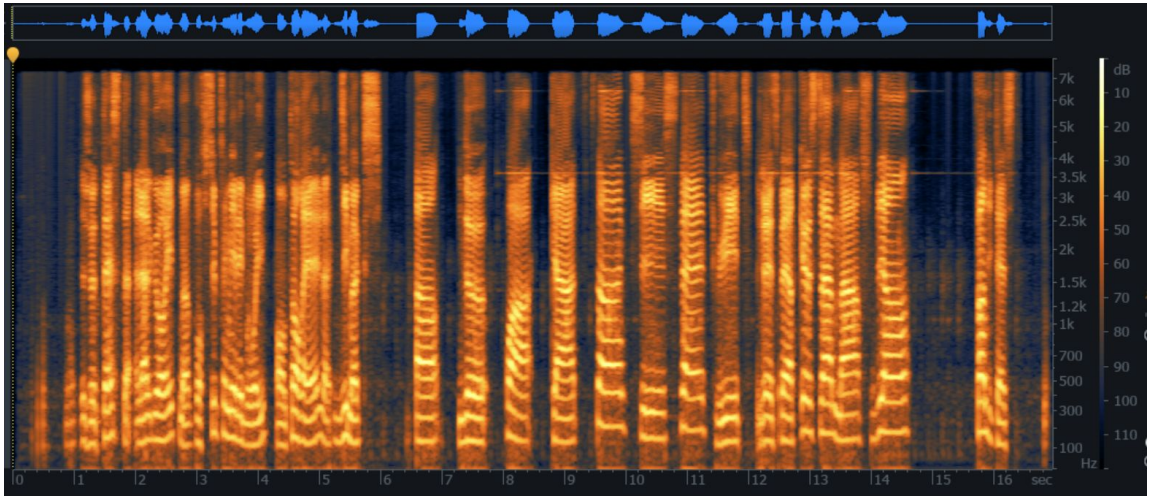
Spectrogram information

Can we use phase?..

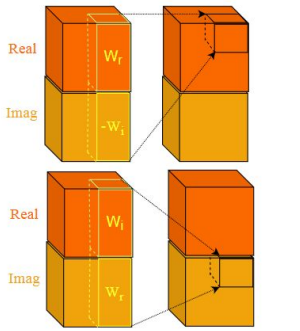


??

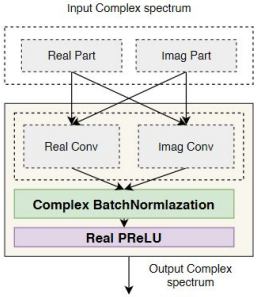
Spectrogram (just Amplitude info)



DCCRN(DNS Challenge, 2020)



(a) complex convolution



(b) complex encoder

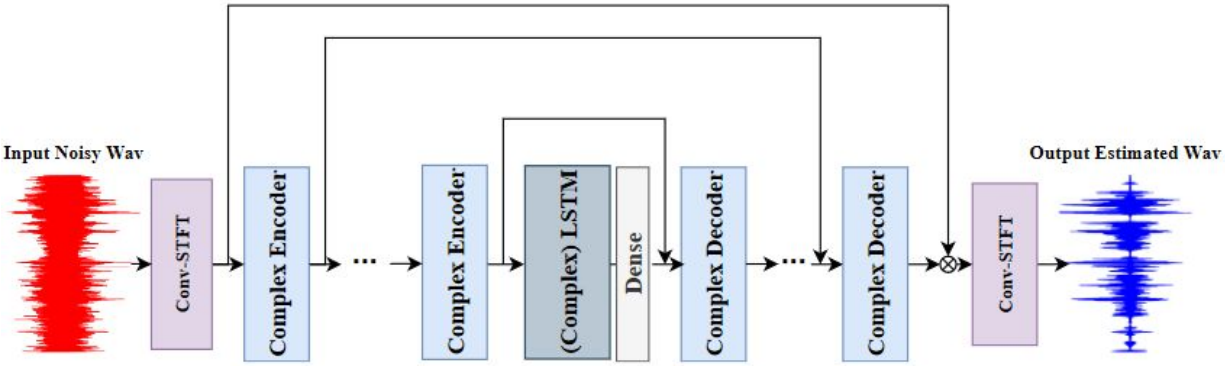


Figure 2: *Complex module*

Ref: DCCRN(2020)

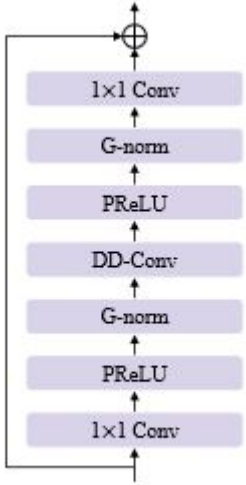
$$F_{out} = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r)$$

Why not to use raw spectrogram (with phases)?
=> complex operations with complex data

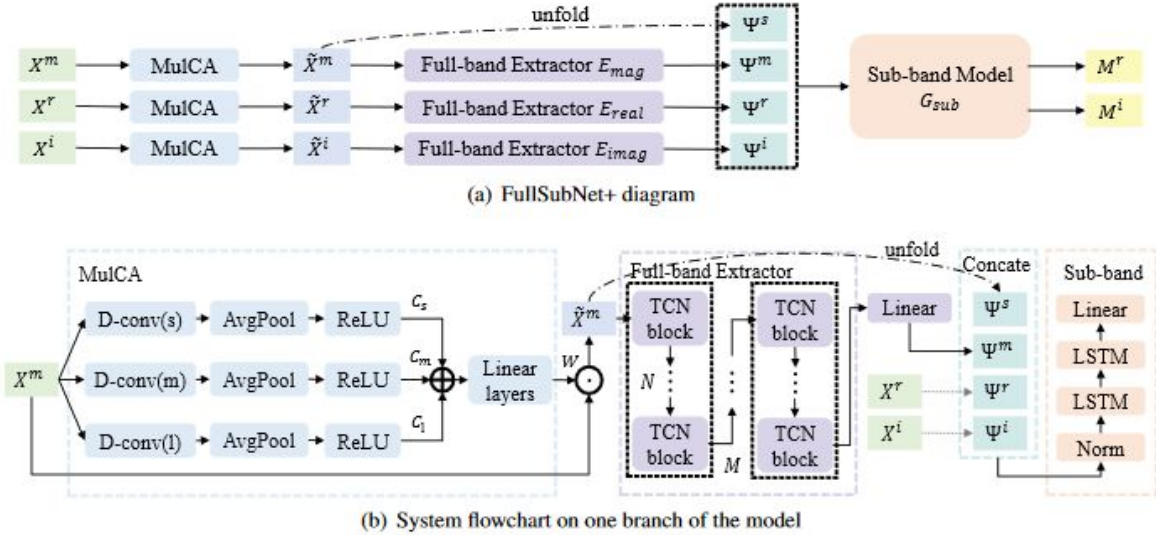
Gives better PESQ scores but a lot of parameters

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

FullSubNet+(2022)



TCN-Block



Idea: use separately magnitude and 2-component phase, encode it via dilated convolutions then fully-convolutional

Ref: FullSubNet+

FullSubNet+(2022)

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**

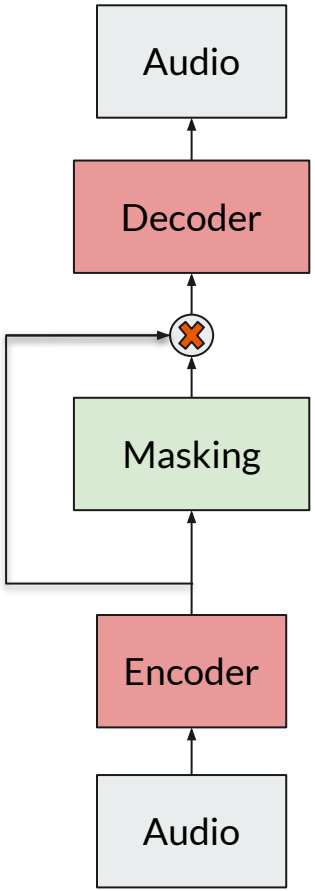
DEMUCS Denoiser (Facebook, 2020)

DEMUCS Architecture

ConvTranspose with skip

2-Layer LSTM (in time)

CNN with skip

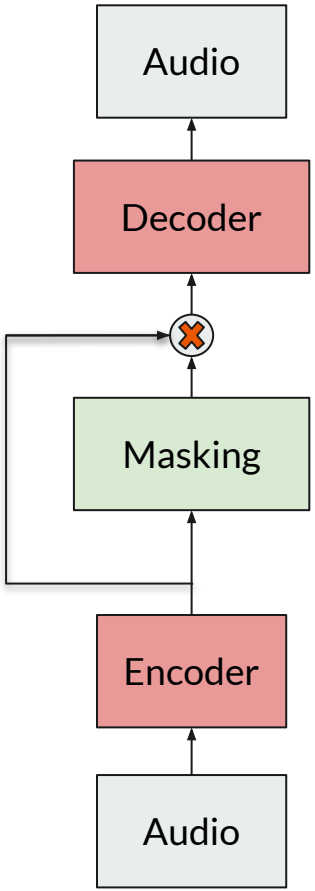


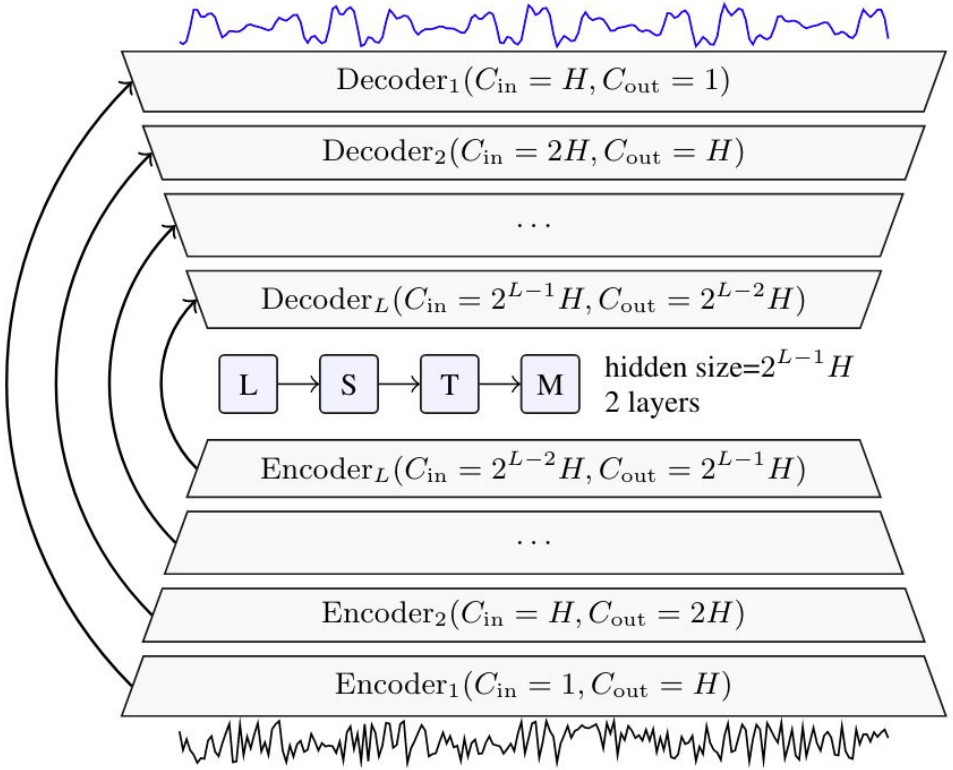
DEMUCS Architecture

ConvTranspose with skip

2-Layer LSTM (in time)

CNN with skip





Ref: DEMUCS(Facebook 2020)

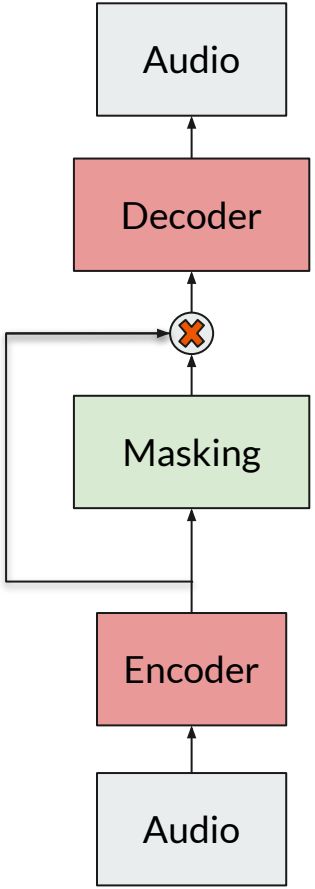
DEMUCS Architecture

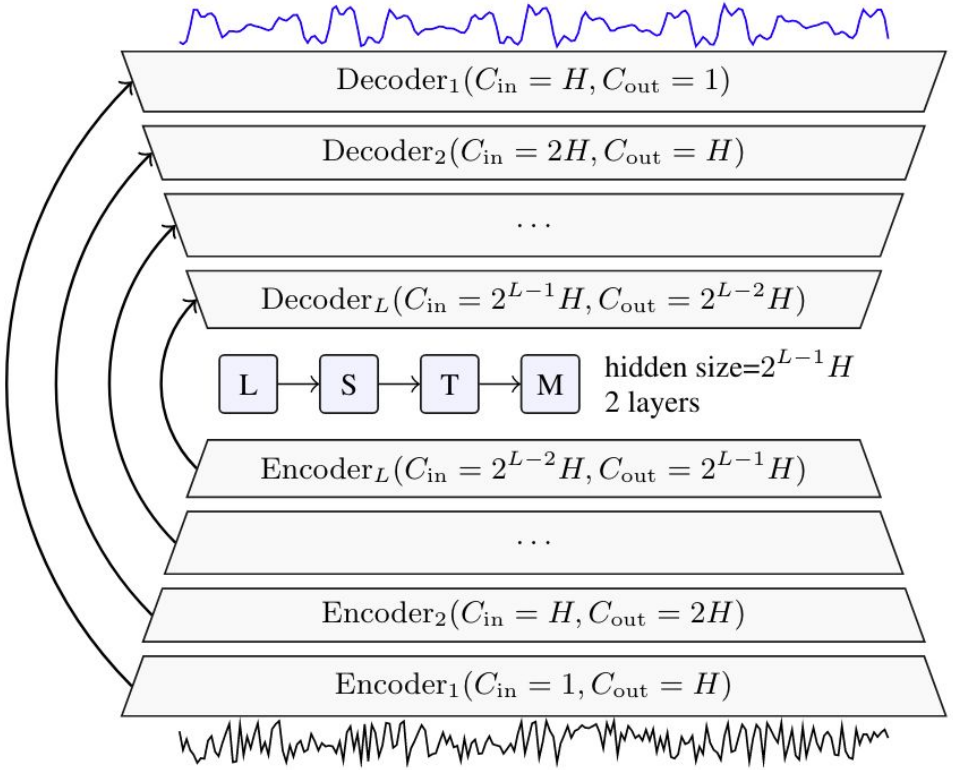
ConvTranspose with skip

2-Layer LSTM (in time)

CNN with skip

UNet?...

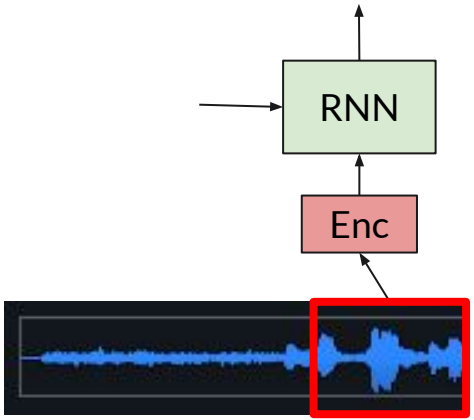




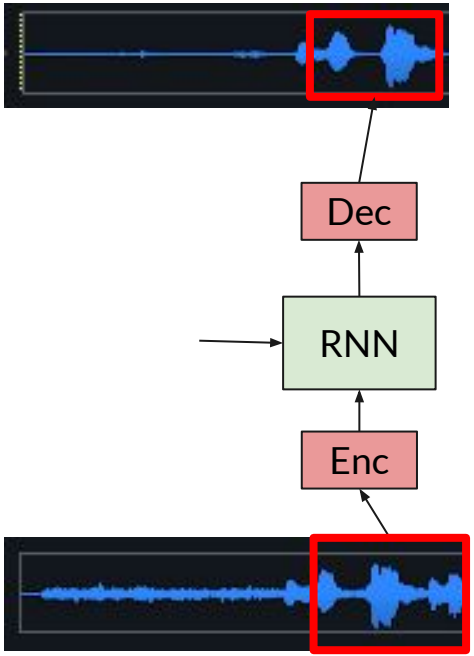
Ref: DEMUCS(Facebook 2020)

DEMUCS Architecture

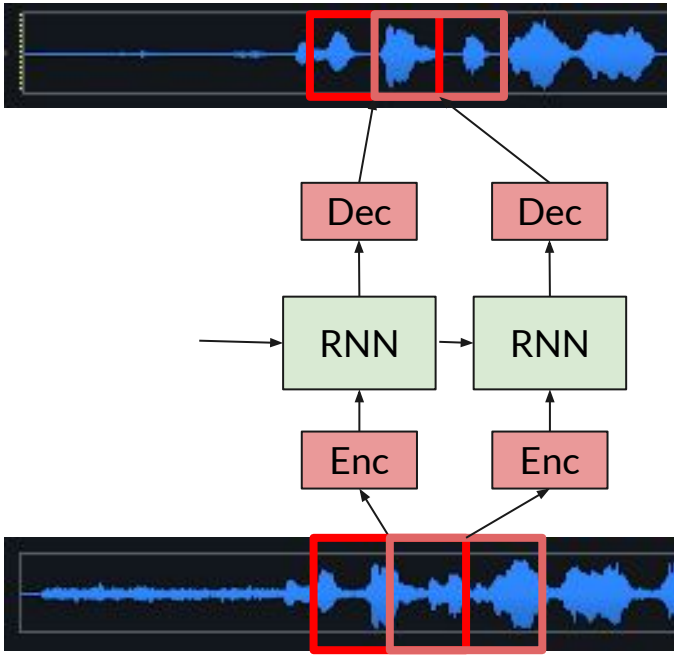
DEMUCS Architecture



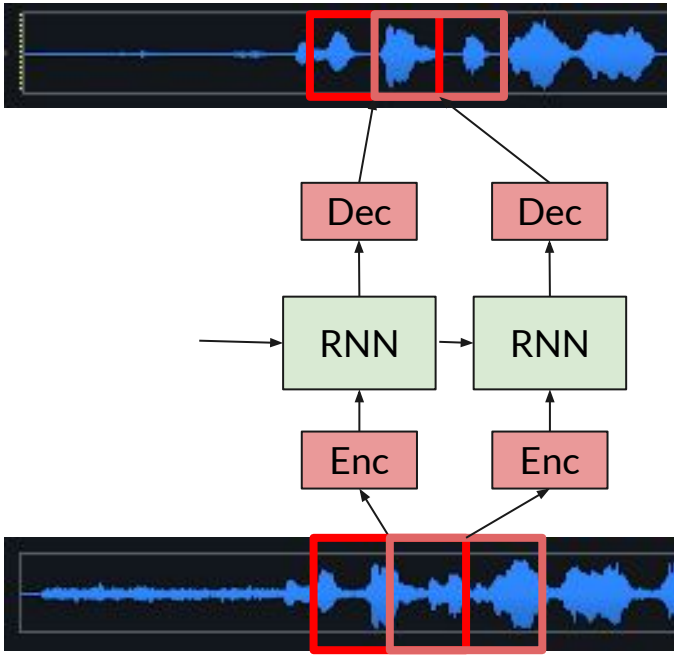
DEMUCS Architecture



DEMUCS Architecture

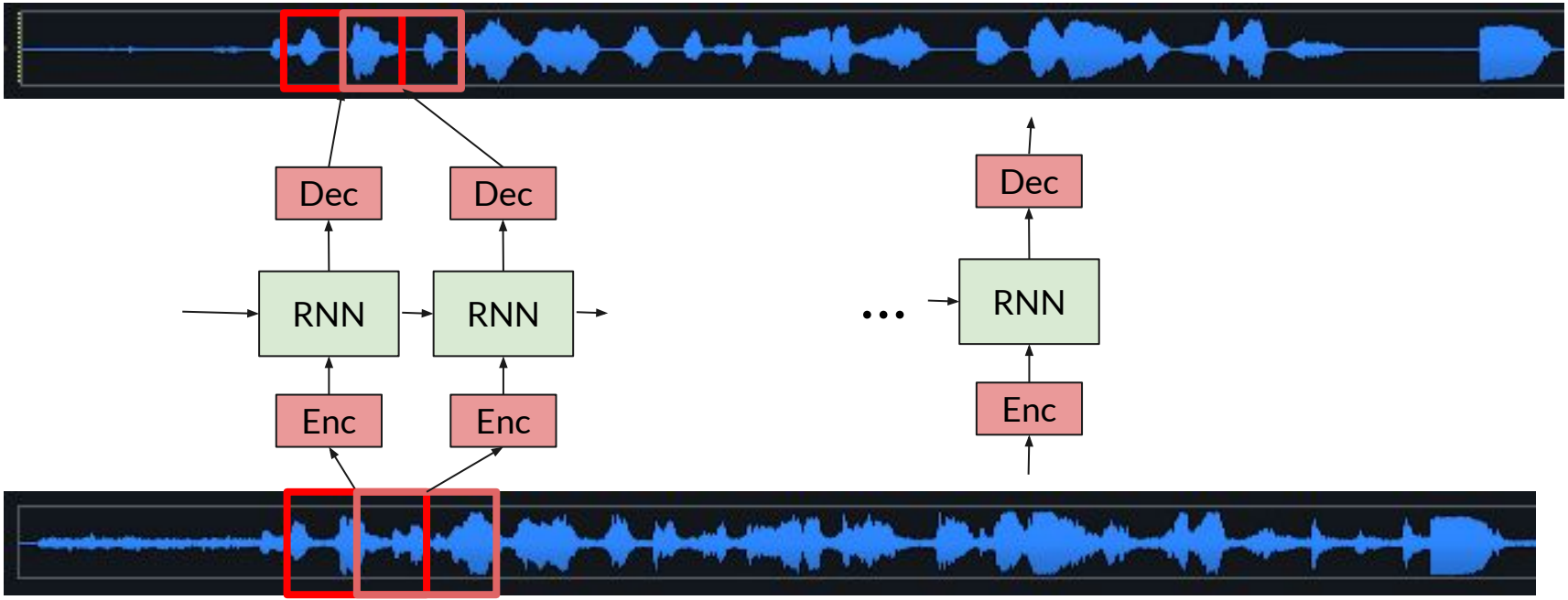


DEMUCS Architecture



...

DEMUCS Architecture



DEMUCS Metrics

Architecture	Wav?	Extra?	Test SDR in dB				
			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

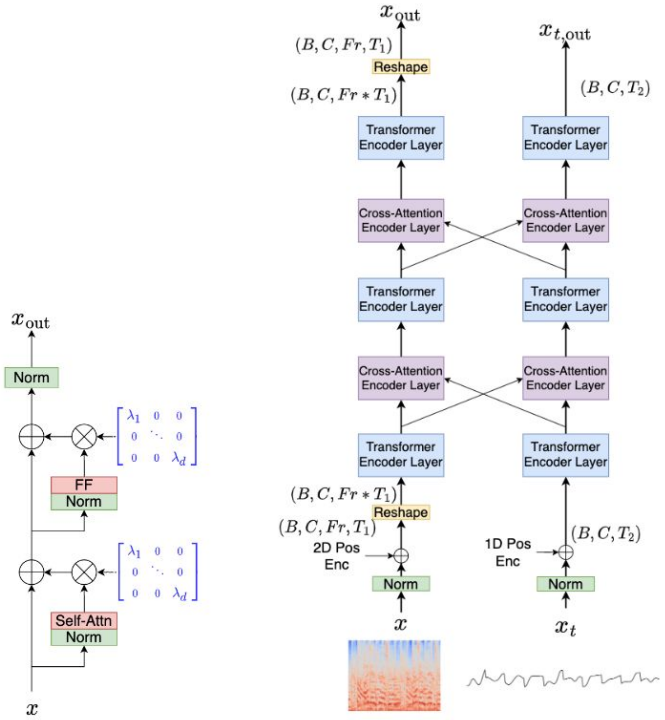
*: each track is only 30 seconds, †: from current work, ††: trained without pitch/tempo augmentation, as it deteriorates performance.

Ref: DEMUCS(2020)

Hybrid Transformer DEMUCS (Facebook, 2022)

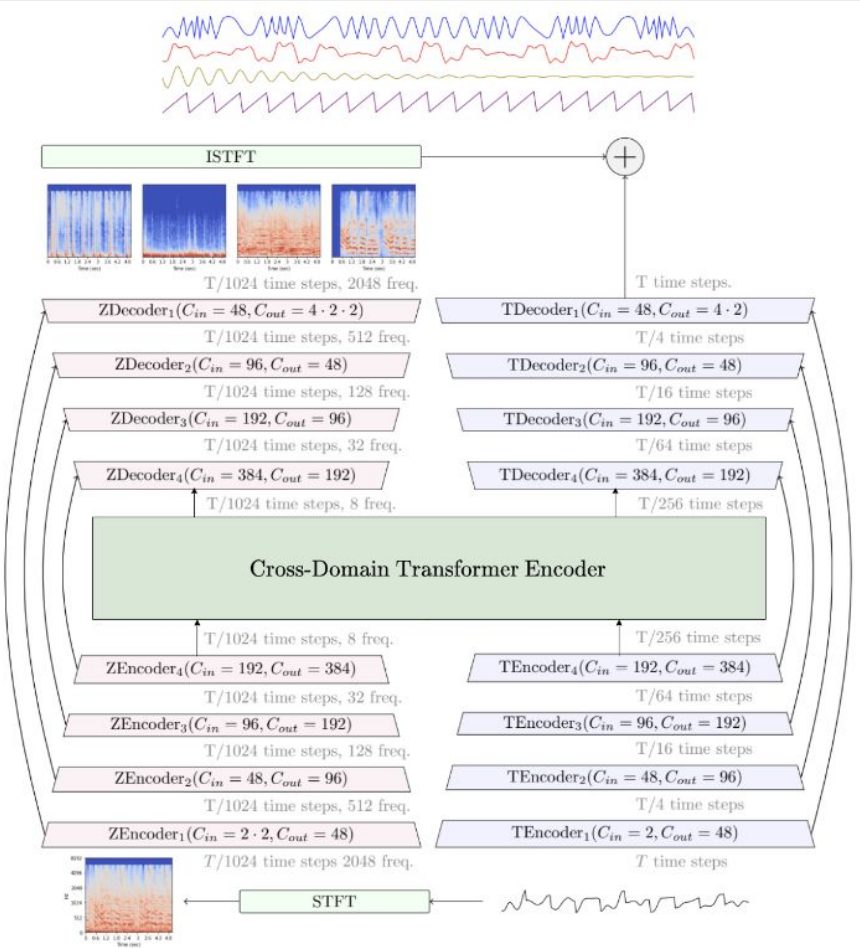
Why not to use both?

Ref: HT DEMUCS(2022)



(a) Transformer Encoder Layer

(b) Cross-domain Transformer Encoder of depth 5



Metrics for HT Demucs

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**

Metrics for HT Demucs

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**

Still the best to separate BASS !

Transforming the sound…

Source Separation literally means separate any source of particular interest…

Speech again but limited to

Separate signal (music, speech…) from noise

Denoising

Separate audio mix into separate tracks

Source Separation

Blind (any k, cocktail)

Guided (specifically bass, vocal, guitar)

Target (specific source given by reference)

Guess when specific sound sources are active

Diarization

Transforming the sound…

Source Separation literally means separate any source of particular interest…

Speech again but limited to

Separate signal (music, speech…) from noise

Denoising

Blind (any k, cocktail)

Separate audio mix into separate tracks

Source Separation

Guided (specifically bass, vocal, guitar)

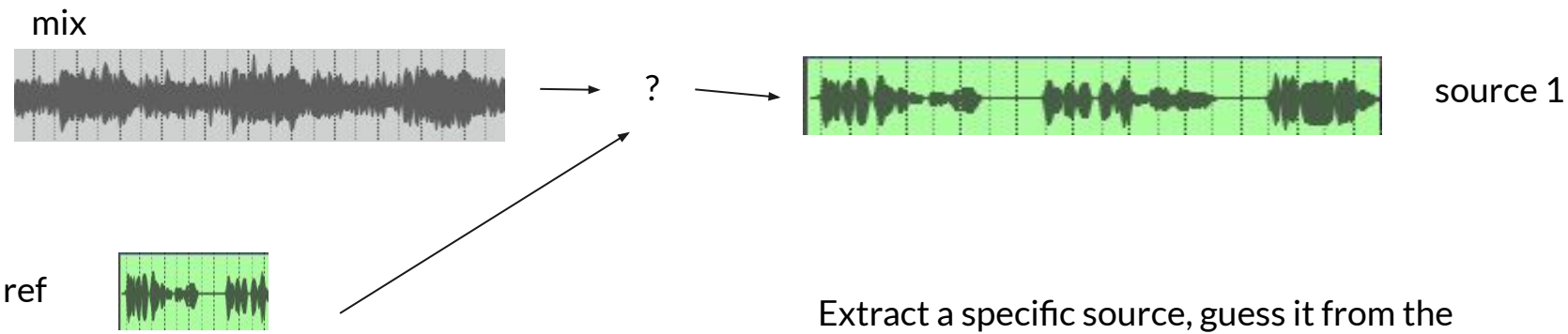
Target (specific source given by reference)

Guess when specific sound sources are active

Diarization

Target Source Separation(TSS) Problem

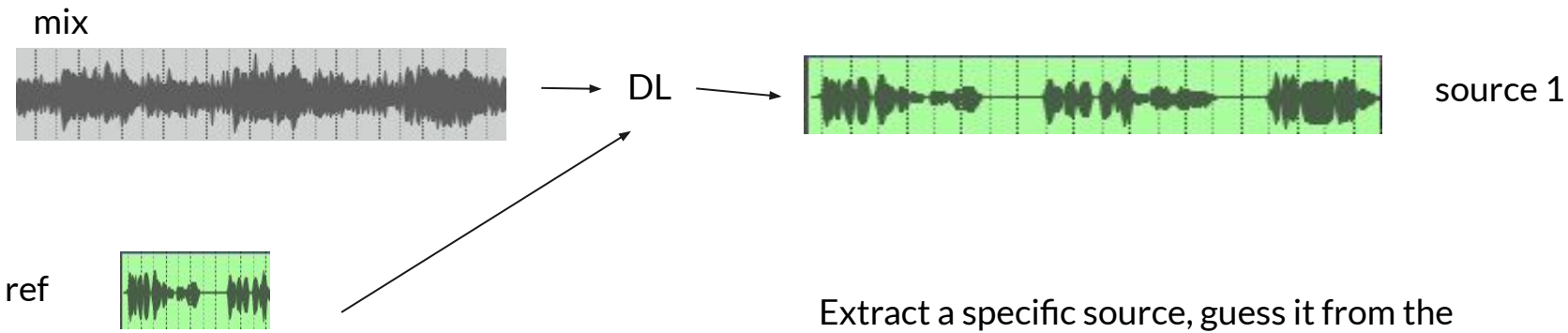
Goal: extract a particular source from the (noised mix)



Extract a specific source, guess it from the given audio reference

Target Source Separation(TSS) Problem

Goal: extract a particular source from the (noised mix)

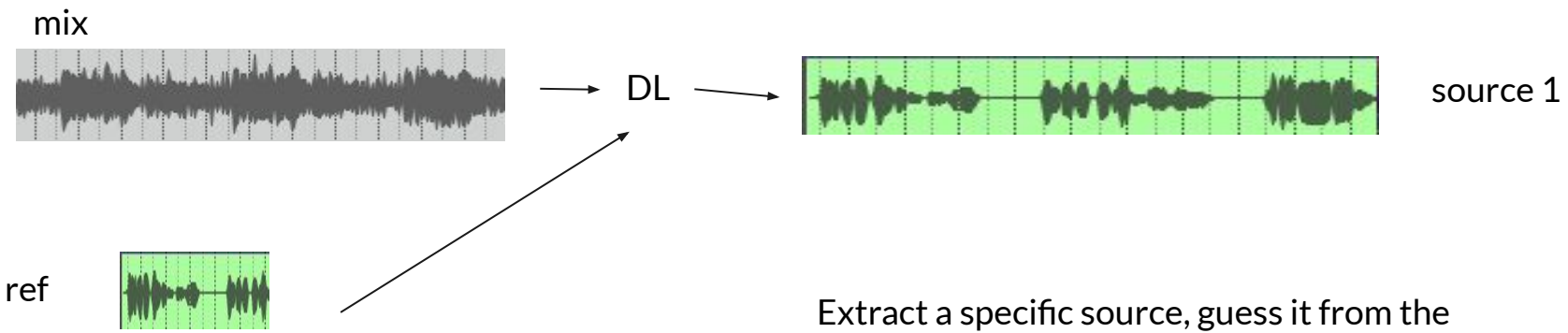


Extract a specific source, guess it from the given audio reference

Application: filter speaker voice in conference call systems of smart devices when in crowded environment

Target Source Separation(TSS) Problem

Goal: extract a particular source from the (noised mix)



Extract a specific source, guess it from the given audio reference

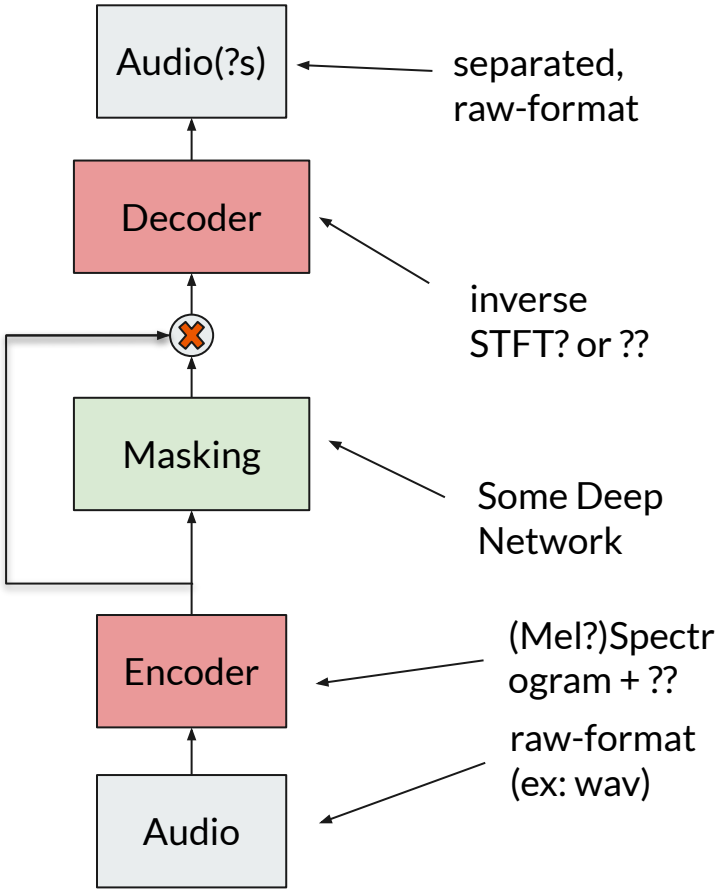
Application: filter speaker voice in conference call systems of smart devices when in crowded environment

Encoder-Separator-Decoder Pipeline (one more time)

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

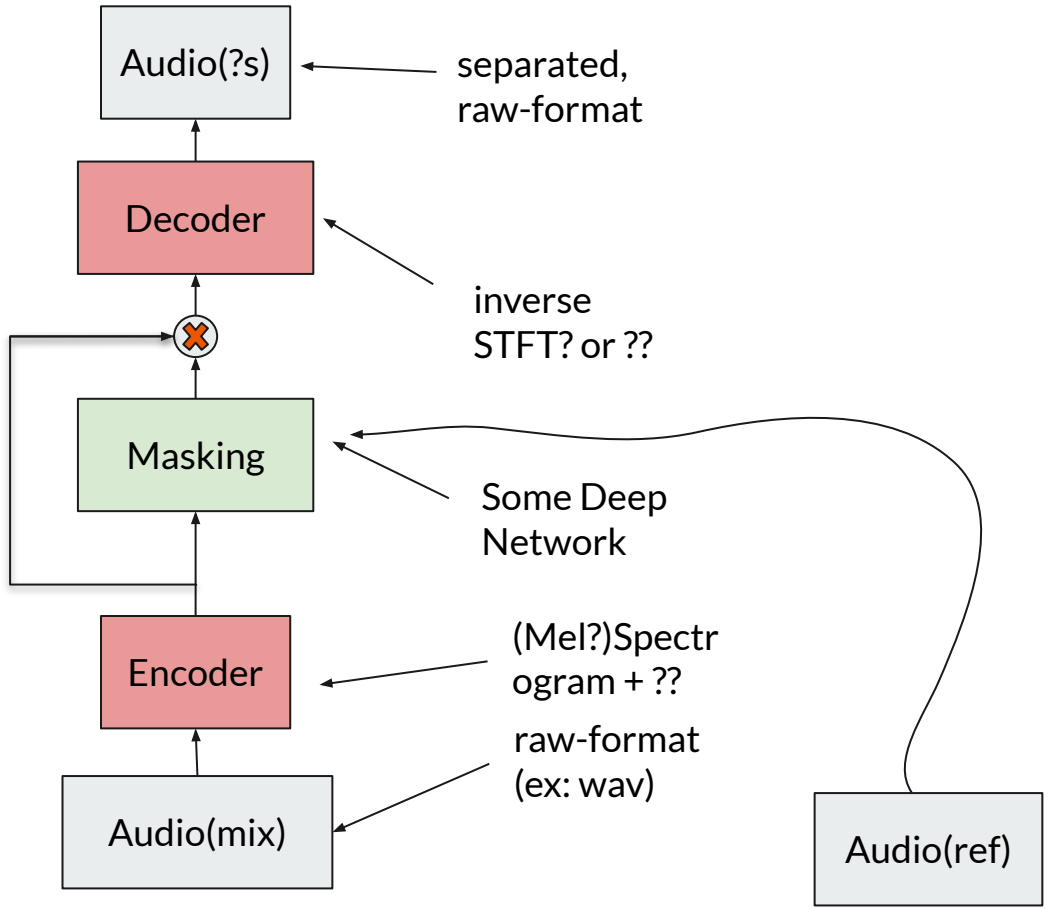


Encoder-Separator-Decoder Pipeline (+reference)

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

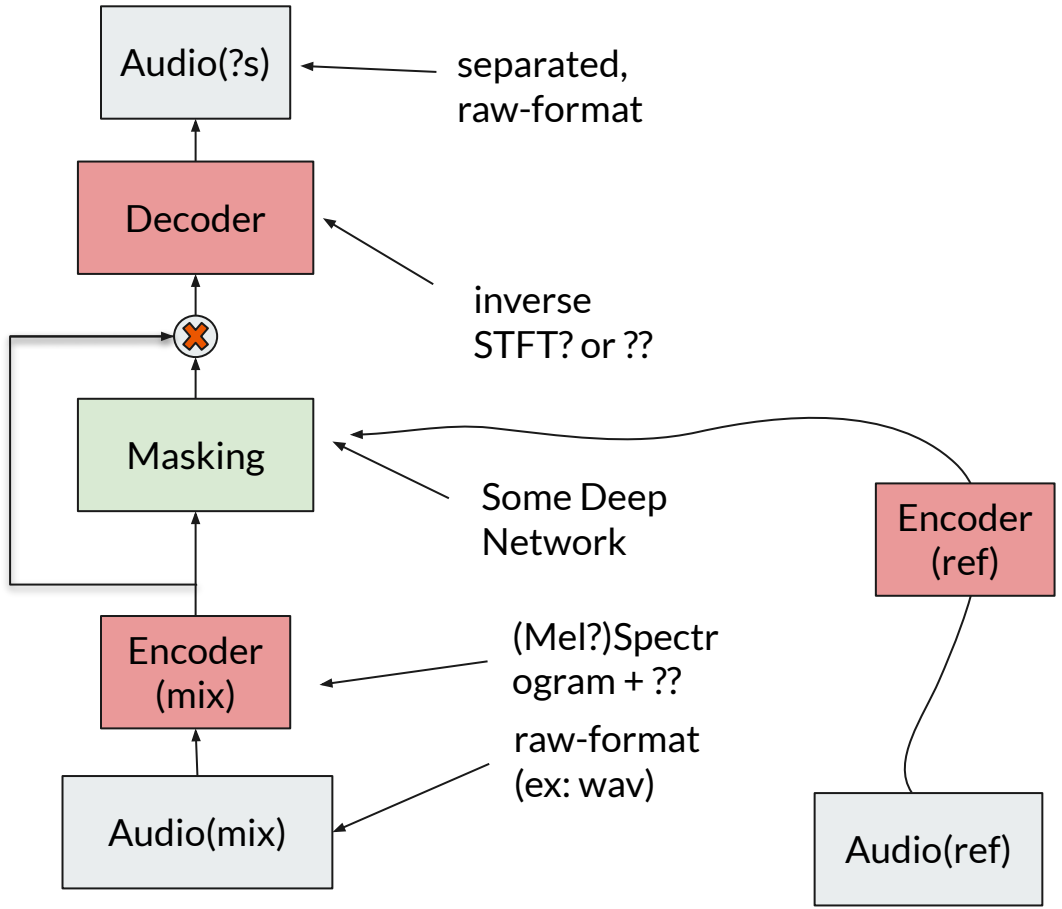


Encoder-Separator-Decoder Pipeline (+reference)

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

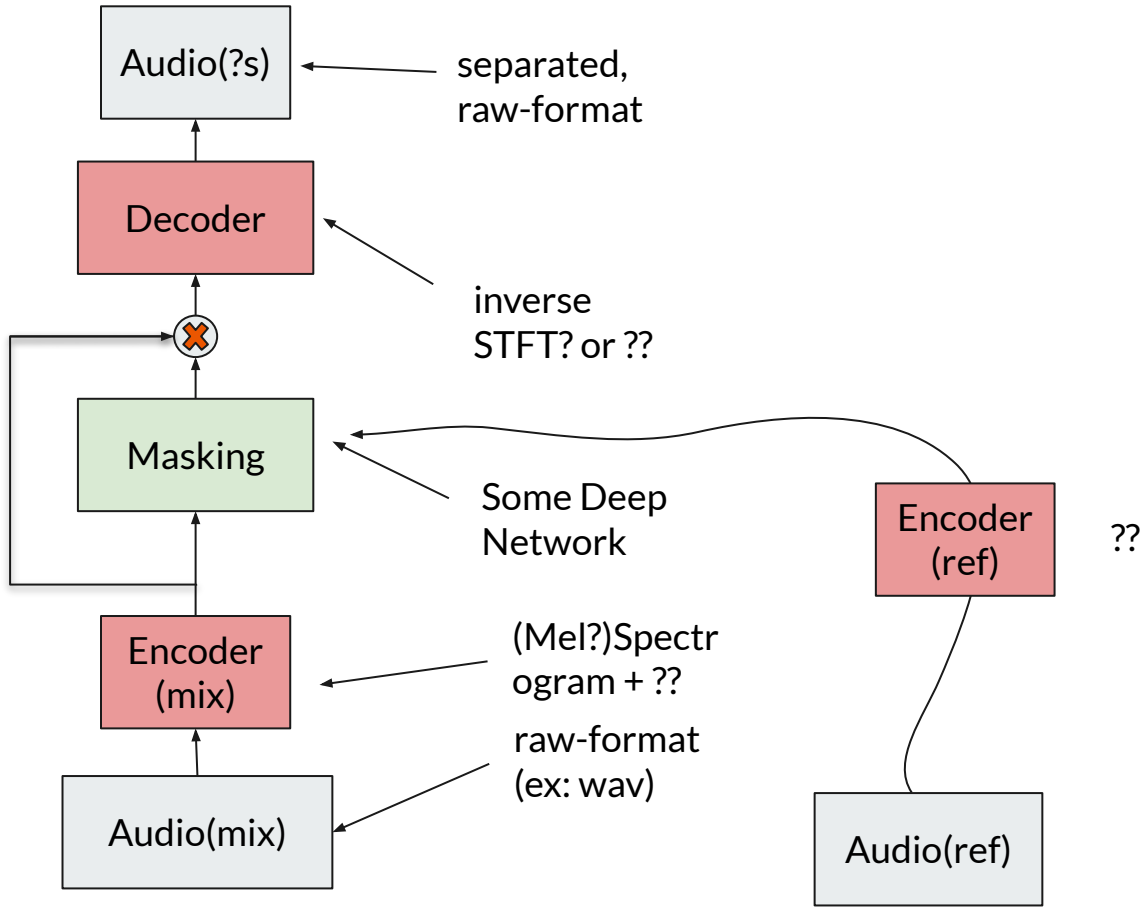


Encoder-Separator-Decoder Pipeline (+reference)

Deep Learning approach: see what can be done on the spectrogram

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

In a nontrivial way…

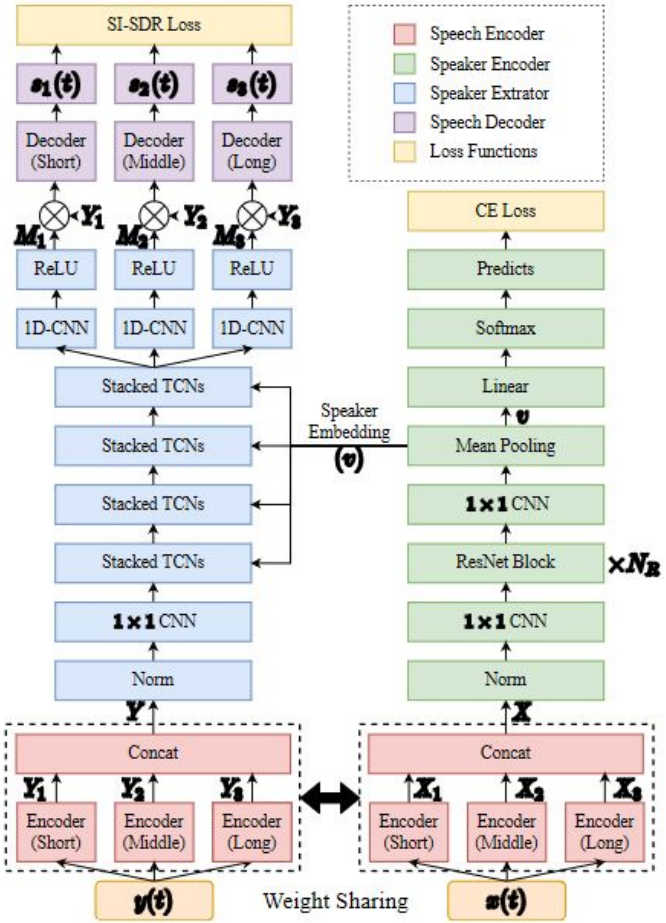


SpEx+ (2020) for TSS

Goal: solve target speaker separation

- **1D 3-component CNN** Encoder
- **TCN ResNet-like structure** as Separator
- **ConvTranspose** decoders
- Arbitrary-length audio as input
- **Classifier** component
- **Same** encoder for mix and ref

Trained and evaluated on WSJ0-2mix



Ref.: SpEx+(2020)

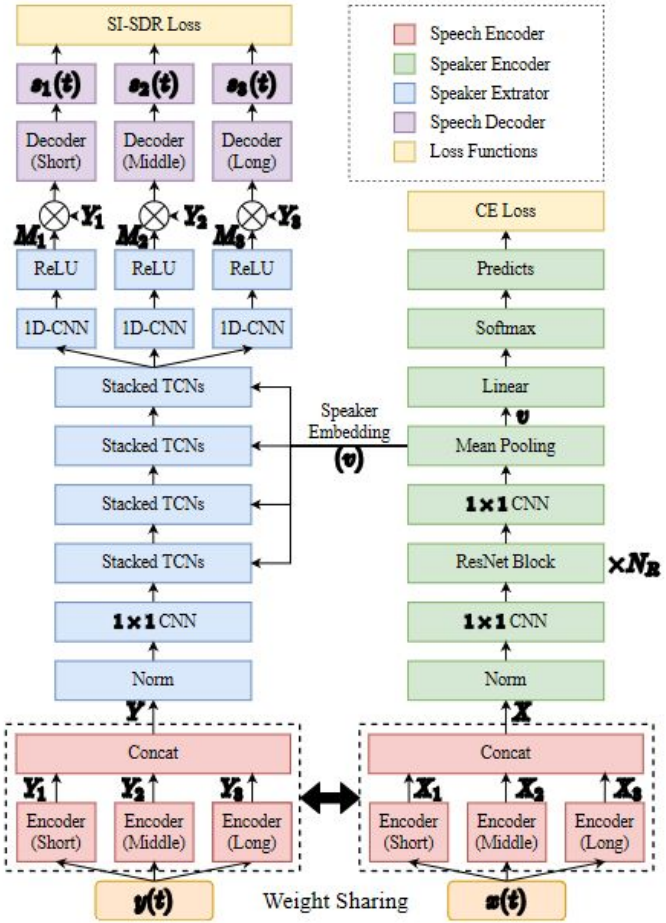
SpEx+ (2020) for TSS

Goal: solve target speaker separation

- **1D 3-component CNN** Encoder
- **TCN ResNet-like structure** as Separator
- **ConvTranspose** decoders
- Arbitrary-length audio as input
- **Classifier** component
- **Same** encoder for mix and ref
- Streaming?...

Trained and evaluated on WSJ0-2mix

Ref.: SpEx+(2020)



SpEx+ (2020) for TSS

Goal: solve target speaker separation

- **1D 3-component CNN** Encoder
- **TCN ResNet-like structure** as Separator
- **ConvTranspose** decoders
- Arbitrary-length audio as input
- **Classifier** component
- **Same** encoder for mix and ref
- Streaming?...

Trained and evaluated on WSJ0-2mix

Task	Methods	#Params	SDRi	SI-SDR
BSS	DPCL++ [5]	13.6M	-	10.8
	uPIT-BLSTM-ST [10]	92.7M	10.0	-
	DANet [7]	9.1M	-	10.5
	cuPIT-Grid-RD [11]	53.2M	10.2	-
	SDC-G-MTL [25]	53.9M	10.5	-
	CBLDNN-GAT [26]	39.5M	11.0	-
	Chimera++ [6]	32.9M	12.0	11.5
	WA-MISI-5 [27]	32.9M	13.1	12.6
BSS	BLSTM-TasNet [12]	23.6M	13.6	13.2
BSS	Conv-TasNet [13]	5.1M	15.6	15.3
BSS	DPRNN-TasNet [28]	2.6M	19.0	18.8
SE	SpEx [14]	10.8M	17.0	16.6
SE	SpEx+	11.1M	17.6	17.4

Ref.: SpEx+(2020)

Materials

[1] DEMUCS (2020-now, with code, papers and metrics)
[2] FullSubNet+ (2022, with code and paper)
[3] DCCRN (2020)
[4] KUIELAB-MDX-Net (2021)
[5] D3Net (2020, multidilated convolutions)
[6] BandSplitRNN (2022)

Some review on what's next...

[1] Dual-Path Transformer Network (DPTN) (2021)
[2] Diffusion-Based DPM-TSE (2023)
[3] USEE (Conditional Diffusion) (2023)
[4] ConvTasNet with Attention (2023)
[5] Fast-Fourier Convolution for Speech Enhancement (2022)

GAN
[6] HiFi++(2023)

LLM-hype-driven
[7] Text-Guided Speech Extraction(2023)
[8] UniAudio(2023)

....