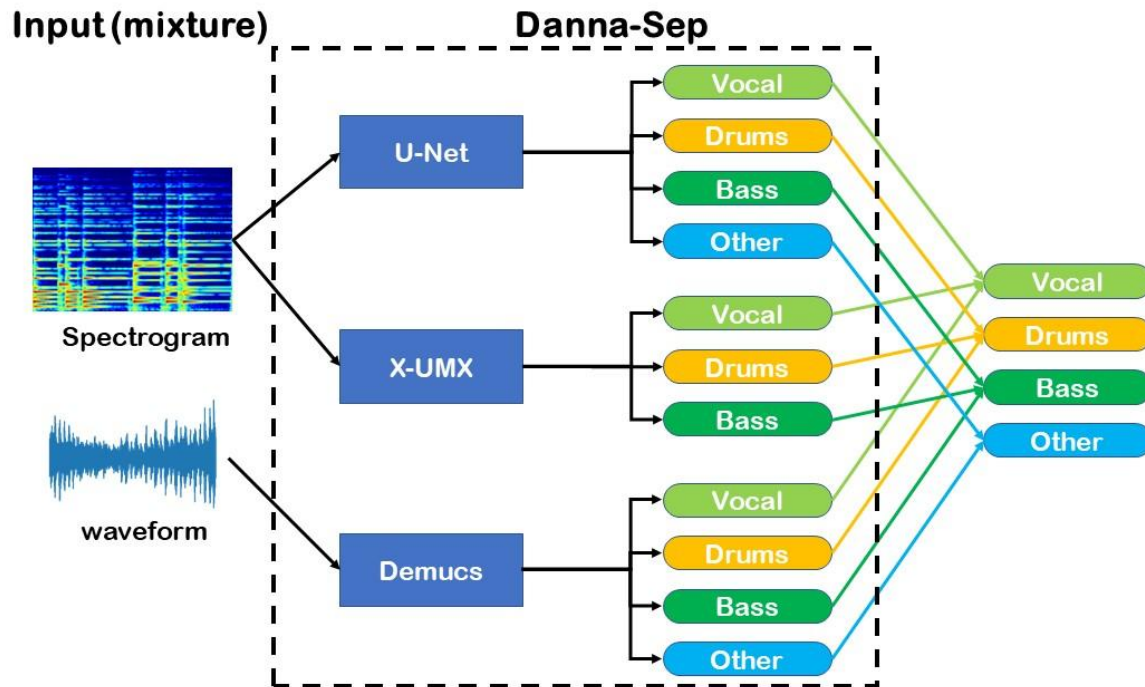


Danna-Sep: Unite to Separate Them All





Chin-Yun Yu, Kin-Wai Cheuk
Music Demixing Workshop, ISMIR 2021

Combining the Best of Frequency and Time



Final rank on MDX

- 4th place on leaderboard A (MUSDB18-HQ)
- 6th place on leaderboard B (extra data is allowed)

Δ	#	Participants	SDR (Song)	SDR (Bass)	SDR (Drums)	SDR (Other)	SDR (Vocals)	External Dataset Used
•	01	 defossez	7.328	8.115	8.037	5.193	7.968	False
•	02	 kuielab	7.236	7.232	7.173	5.636	8.901	False
•	03	 Music_AI	6.882	7.273	7.371	5.091	7.792	False
•	04	 Kazane_Ryo_n...	6.649	6.993	7.018	4.901	7.686	False

Model Fusion

- Blending outputs from three different model linearly

$$\mathcal{S} = w_1 \hat{\mathcal{S}}_1 + w_2 \hat{\mathcal{S}}_2 + w_3 \hat{\mathcal{S}}_3$$

	Drums	Bass	Other	Vocals
Model 1 (frequency masking)	0.2	0.2	0	0.2
Model 2 (frequency masking)	0.2	0.17	0.5	0.4
Model 3 (time domain)	0.6	0.73	0.5	0.4

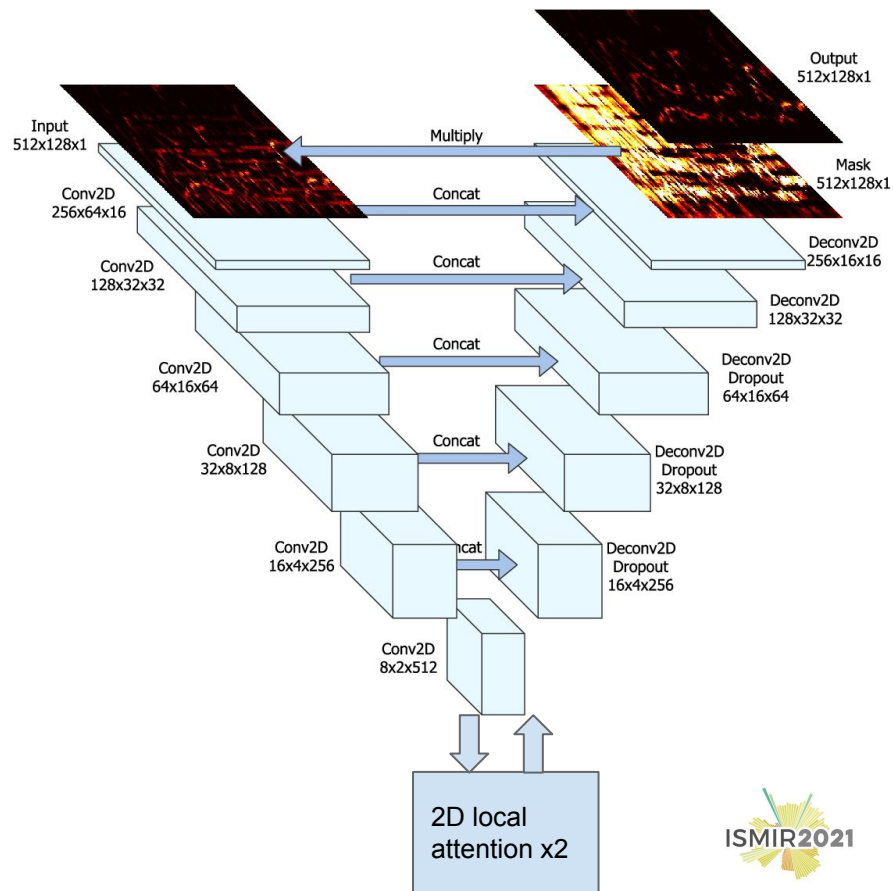
Model 1: Improving X-UMX

- Two tiny tweaks on the loss function
 - Calculate the frequency-domain mse loss using complex value instead of absolute value
 - Apply one iteration of Multichannel Wiener Filtering (MWF) before calculating the loss
 - Our differentiable norbert package: [yoyololicon/norbert: Painless Wiener filters for audio separation \(github.com\)](https://github.com/yoyololicon/norbert)
 - Improvements: complex mse + MWF > real mse + MWF >>>> complex mse

$$\mathcal{L}_{MSE} = \sum_{t,f} \{|Y(t, f)| - |\hat{Y}(t, f)|\}^2$$
$$\Rightarrow \mathcal{L}_{MSE} = \sum_{t,f} \|Y(t, f) - \hat{Y}(t, f)\|_2^2$$

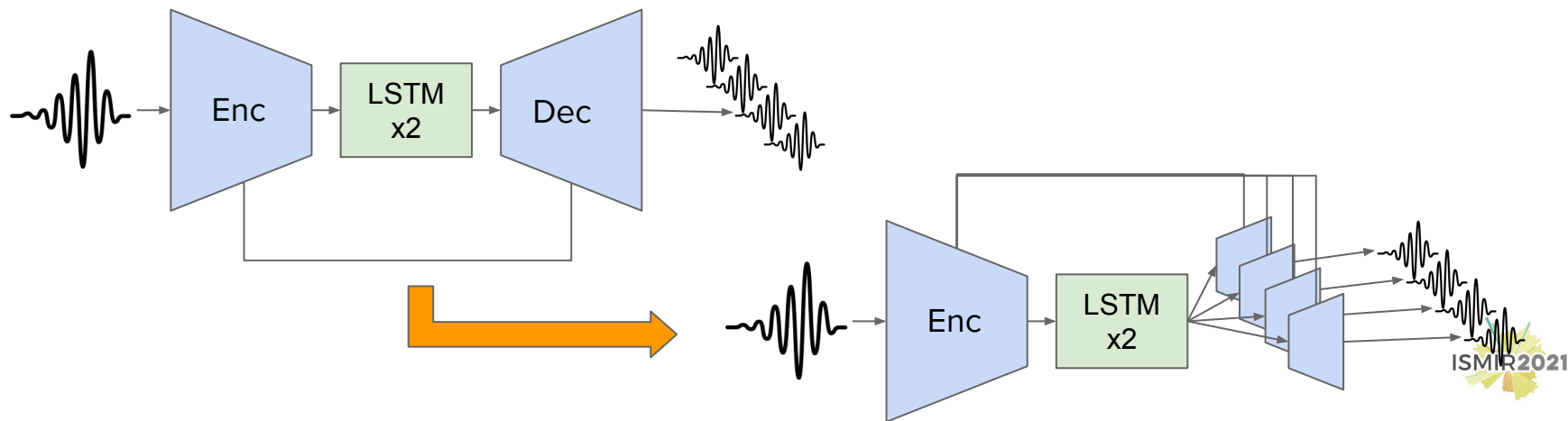
Model 2: U-Net

- Adopt the U-Net structure from **Spleeter**
- Replace each convolution block with a **D3 Block** from **D3Net**
- Add **two local attention layers** as the bottleneck layer
- We also experimented with biaxial biLSTMs (time and frequency axes)



Model 3: Improving Demus

- Using 4 independent decoders, each of which corresponds to one source
 - We adjust the channel size of the decoders to have a comparable number of parameters compared to the original version



Results

SDR score on the test set of MUSDB18

	Drums	Bass	Other	Vocals	Average
X-UMX(baseline)	6.44	5.54	4.46	6.54	5.75
X-UMX(ours)	6.71	5.79	4.63	6.93	6.02
U-Net(ours)	6.43	5.35	4.67	7.05	5.87
Demucs(baseline)	6.67	6.98	4.33	6.89	6.21
Demucs(ours)	6.72	6.97	4.40	6.88	6.24
Danna-Sep	7.20	7.05	5.20	7.63	6.77

Takeaway

- Complex MSE and a more end-to-end training process can gain improvements on X-UMX.
- The encoder part of Demucs is sufficient to separate the necessary info of each sources.
- Even a simple linear combination can effectively combine the complementary strengths of spectrogram models and waveform models.

Acknowledgement

- Sung-Lin Yeh
- Showmin Wang
- Yu-Te Wu
- Yin-Jyun Luo