Decoupling magnitude and phase estimation with deep ResUNet for music source separation

ISMIR 2021 paper overview[1]

Sevag Hanssian

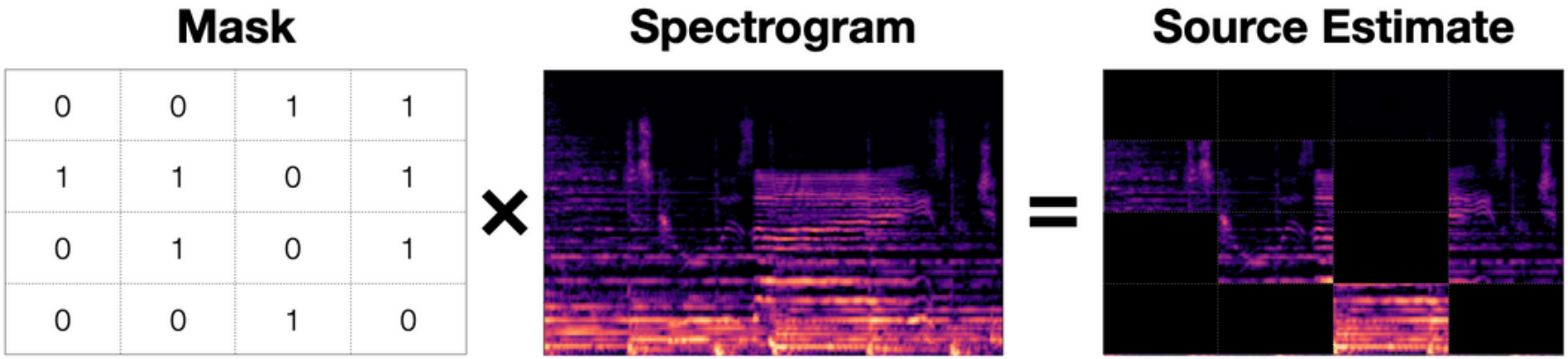
November 4, 2021

[1] Qiuqiang Kong et al. 2021.

Magnitude mask above 1

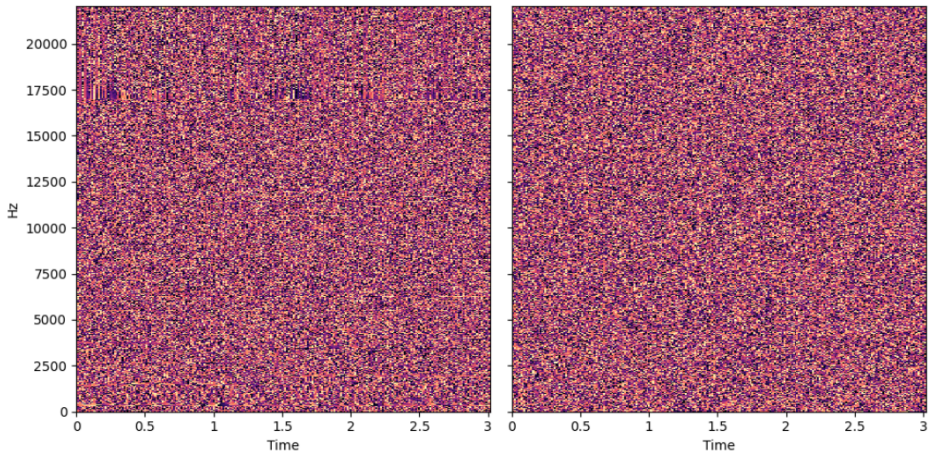
Common approaches to music source separation (MSS):

1. Get spectrogram of mix
2. Take magnitude
3. Multiply by a mask $\in [0, 1]$ to get source estimate
4. Why $[0, 1]$? DFT/STFT is a linear operation:
   $x_a = x_b + x_c, |X_a| = |X_b| + |X_c|$
   $|X_b| = M_b (\in [0, 1]) \times |X_a|$
   if M_b (i.e., Mask of source b) > 1, then $|X_b| > |X_a|$?



Phase!

Common approaches to MSS discard the phase; it's difficult to learn relationships from phase



This paper considers the phase, and uses a complex mask to estimate the magnitude and phase of the spectrogram

$|X_b| = M_b(\in [0,1]) \times |X_a|$

if M_b (i.e., Mask of source b) > 1, then $|X_b| > |X_a|$? **Yes!**

|M(t,f)| can be larger than 1... this may happen when S(t,f) and N(t,f) are out of phase, since that makes the magnitude of mixture to be smaller than that of (individual) signal

Results

Table 2. Comparison of SDRs of previous and our proposed MSS systems.

	vocals	bass	drums	other	acc.
Open-Unmix [14]	6.32	5.23	5.73	4.02	-
Wave-U-Net [22]	3.25	3.21	4.22	2.25	-
Demucs [21]	6.29	5.83	6.08	4.12	-
Conv-TasNet [19]	6.81	5.66	6.08	4.37	-
Spleeter [16]	6.86	5.51	6.71	4.55	
D3Net [35]	7.24	5.25	**7.01**	4.53	13.52
ResUNetDecouple+	**8.98**	**6.04**	6.62	**5.29**	**16.63**

1. We showed that previous MSS methods have upper bound of the performance due to a strong assumption on the magnitude of the masks
2. We also showed that accurate phase estimation and unbound complex ideal ratio masks (cIRMs) are important for MSS
3. Finally, we analyzed the distribution of cRIMs for MSS and showed that 22% of cIRMs have magnitude larger than one