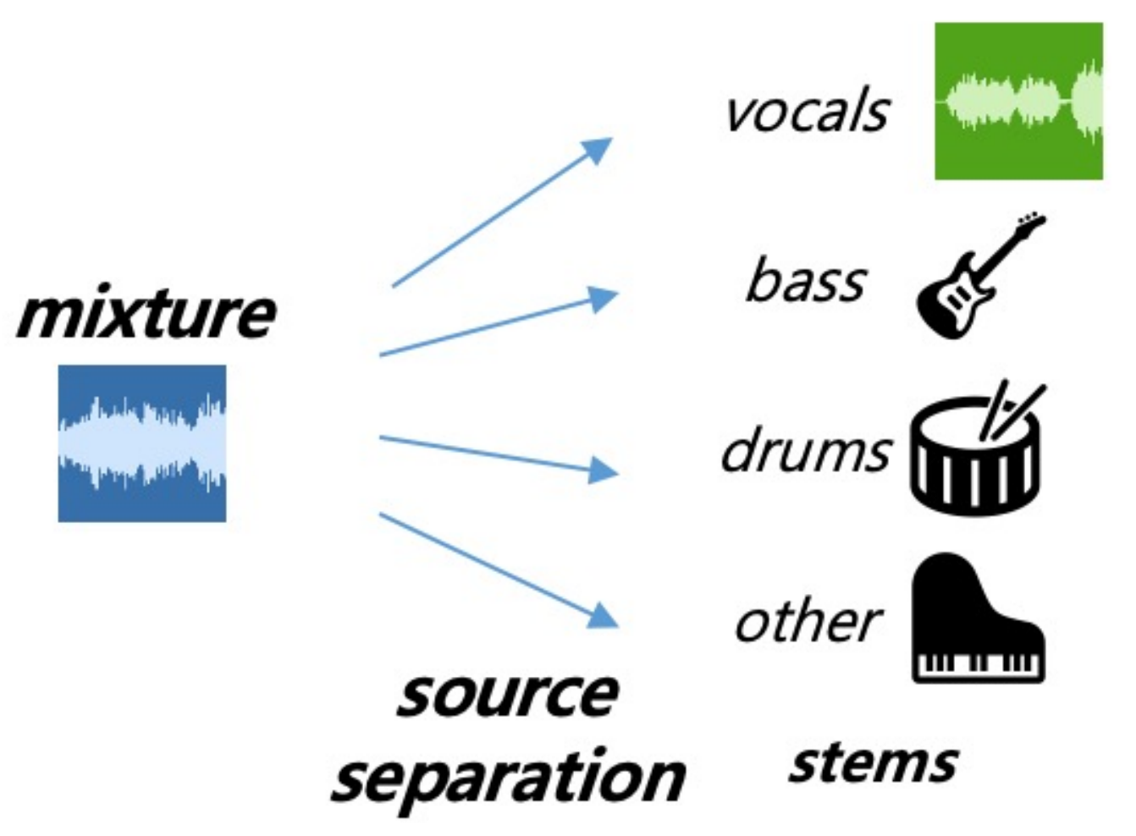
Towards Robust Music Source Separation on Loud Commercial Music

Chang-Bin Jeon and Kyogu Lee
Seoul National University, Music and Audio Research Group

musdb-XL : github.com/jeonchangbin49/musdb-XL
LimitAug : github.com/jeonchangbin49/LimitAug
paper : arxiv.org/abs/2208.14355
contact : vinyne@snu.ac.kr

Music Source Separation

· A task of isolating individual instrumental sources (stems) from music.

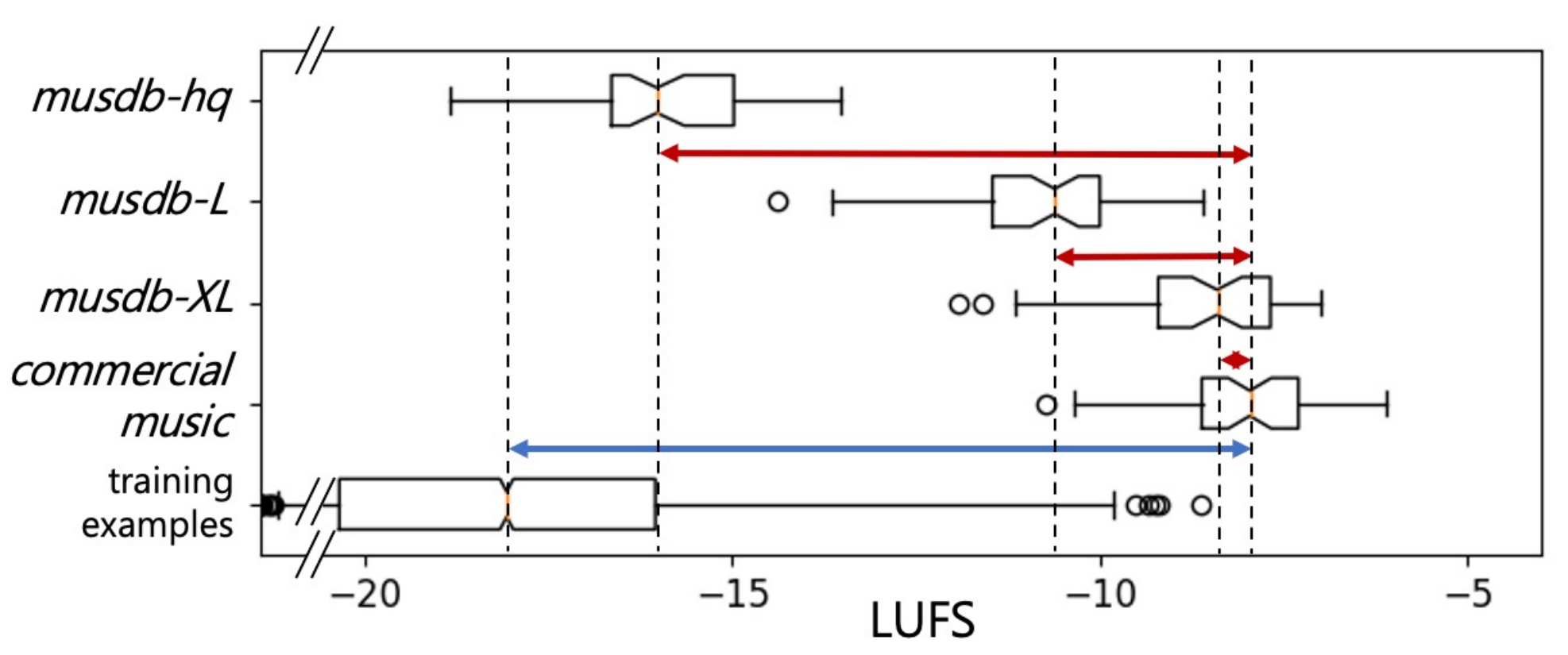


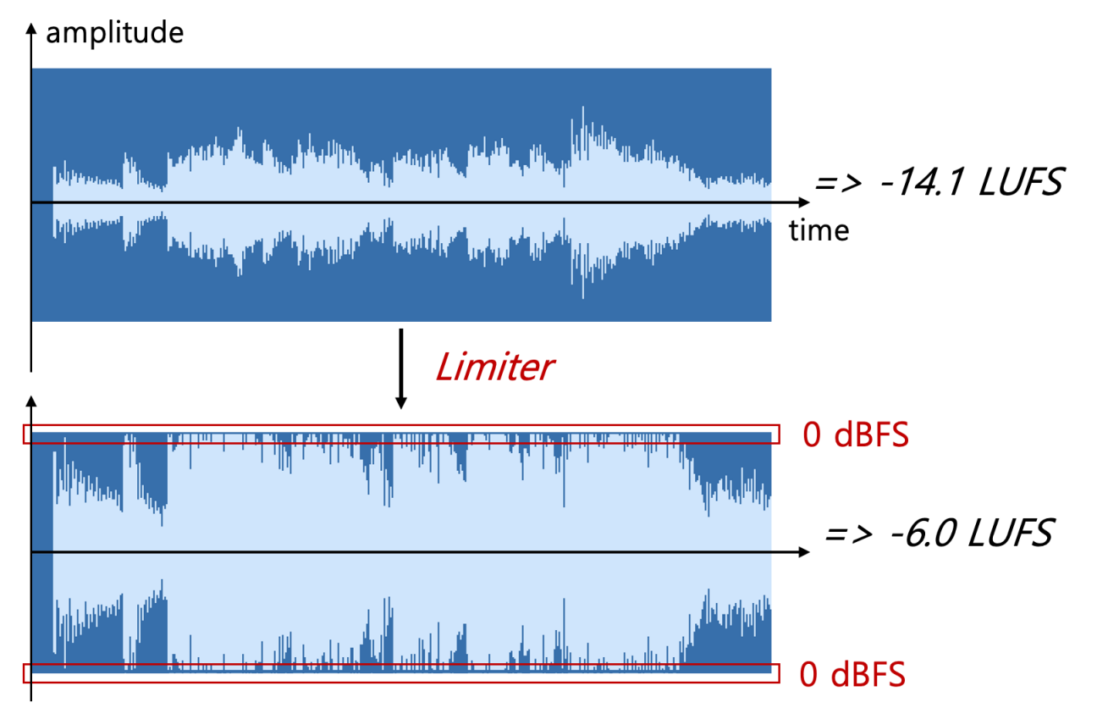
Motivation

- Commercial music has **extreme loudness** and **heavily compressed dynamic range**
  - => Not considered in music source separation yet.
  - => Huge domain shift occurs between train domain and real world.

=> Will the domain shift result in actual performance decrease?

=> If it does, let's decrease the domain shift!





Contributions

1) ***New musdb-XL evaluation data***
   - We introduce *musdb-L* and *musdb-XL* evaluation datasets, which have comparable overall loudness to commercial music, for the evaluation of music source separation.

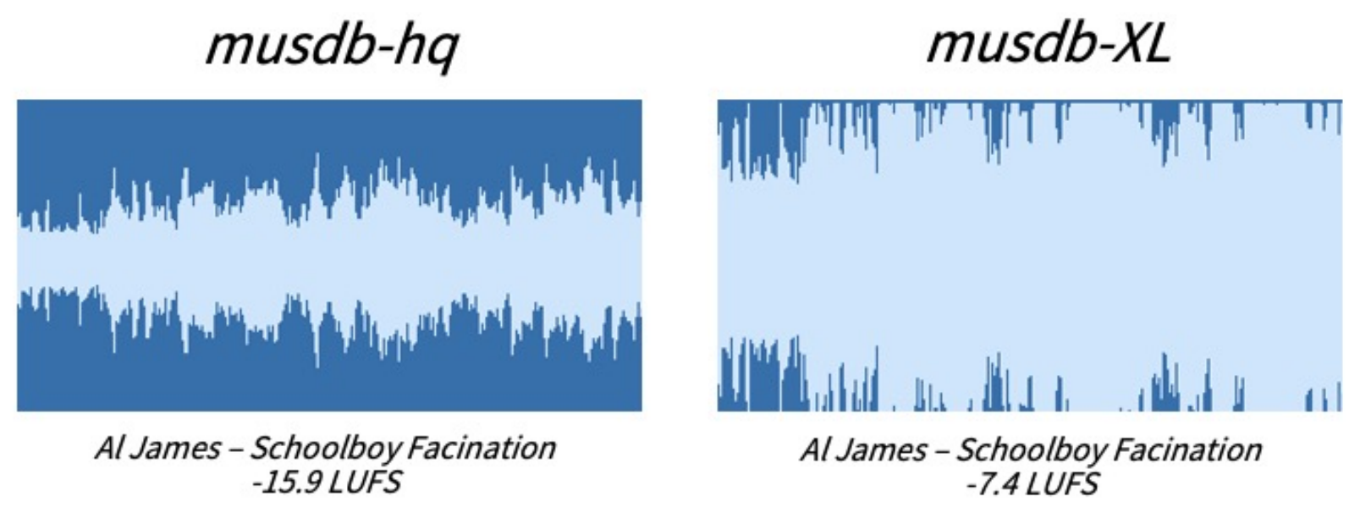
2) ***The domain shift => Actual performance degradation***
   - Using *musdb-L* and *XL*, we quantitatively confirm that the domain shift causes performance degradation of the state-of-the-art networks that were trained without considering loud and compressed music characteristics.

3) ***LimitAug data augmentation***
   - We propose *LimitAug* data augmentation method and experimentally confirm that it is beneficial to alleviate the domain shift between train data and the *musdb-L* or *XL*.

Musdb-XL

- We manually made loud and compressed version of *musdb-hq => musb-L and musdb-XL*

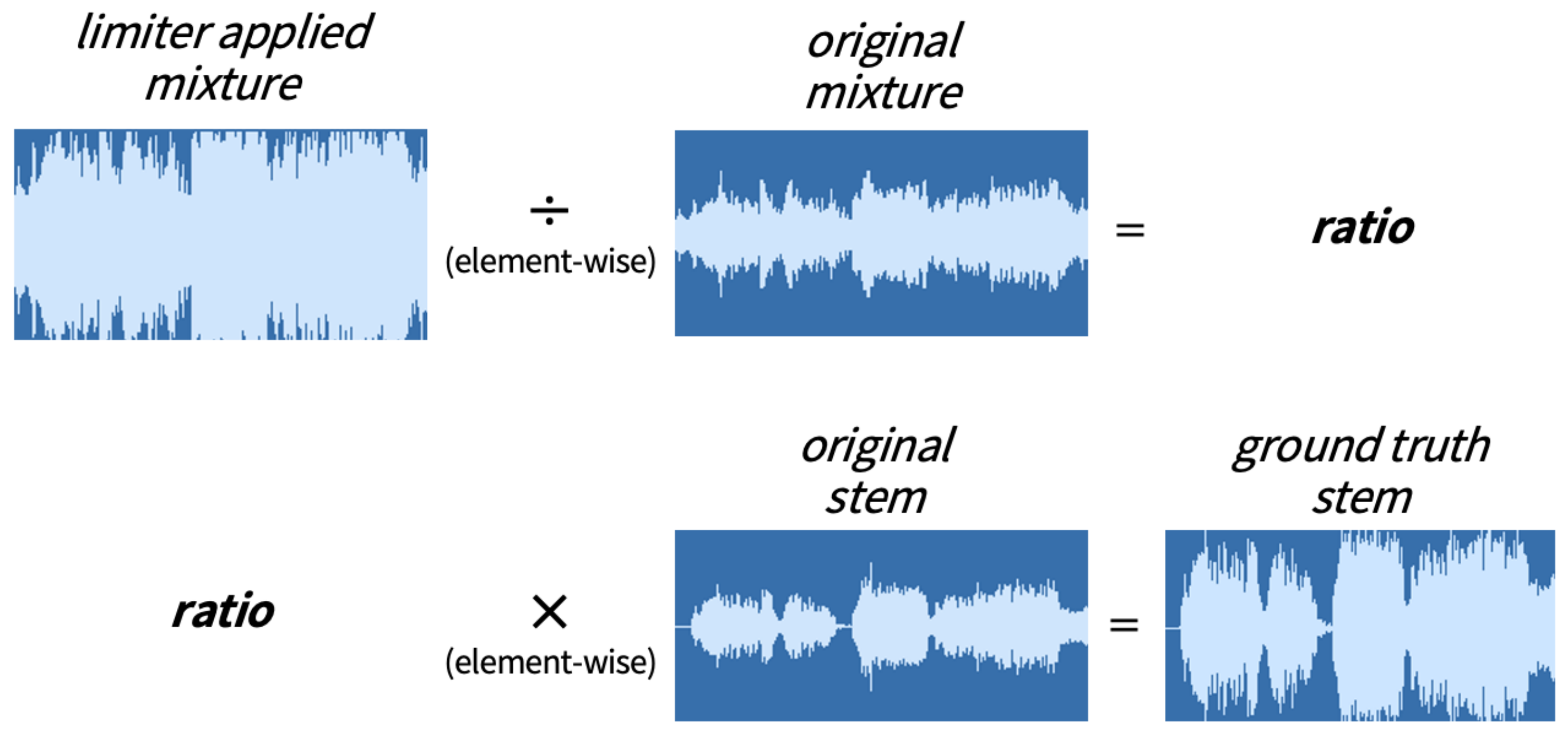
dataset	Loudness [LUFS]			
	min	max	median	mean (std)
musdb-hq	-18.84	-13.52	-16.02	-15.92 (1.27)
musdb-L	-14.39	-8.61	-10.61	-10.89 (1.19)
musdb-XL	-11.93	-6.99	-8.41	-8.61 (1.17)
commercial	**-10.75**	**-6.10**	**-7.96**	**-8.05 (1.06)**



How to get ground truth stems of a limiter applied mixture?

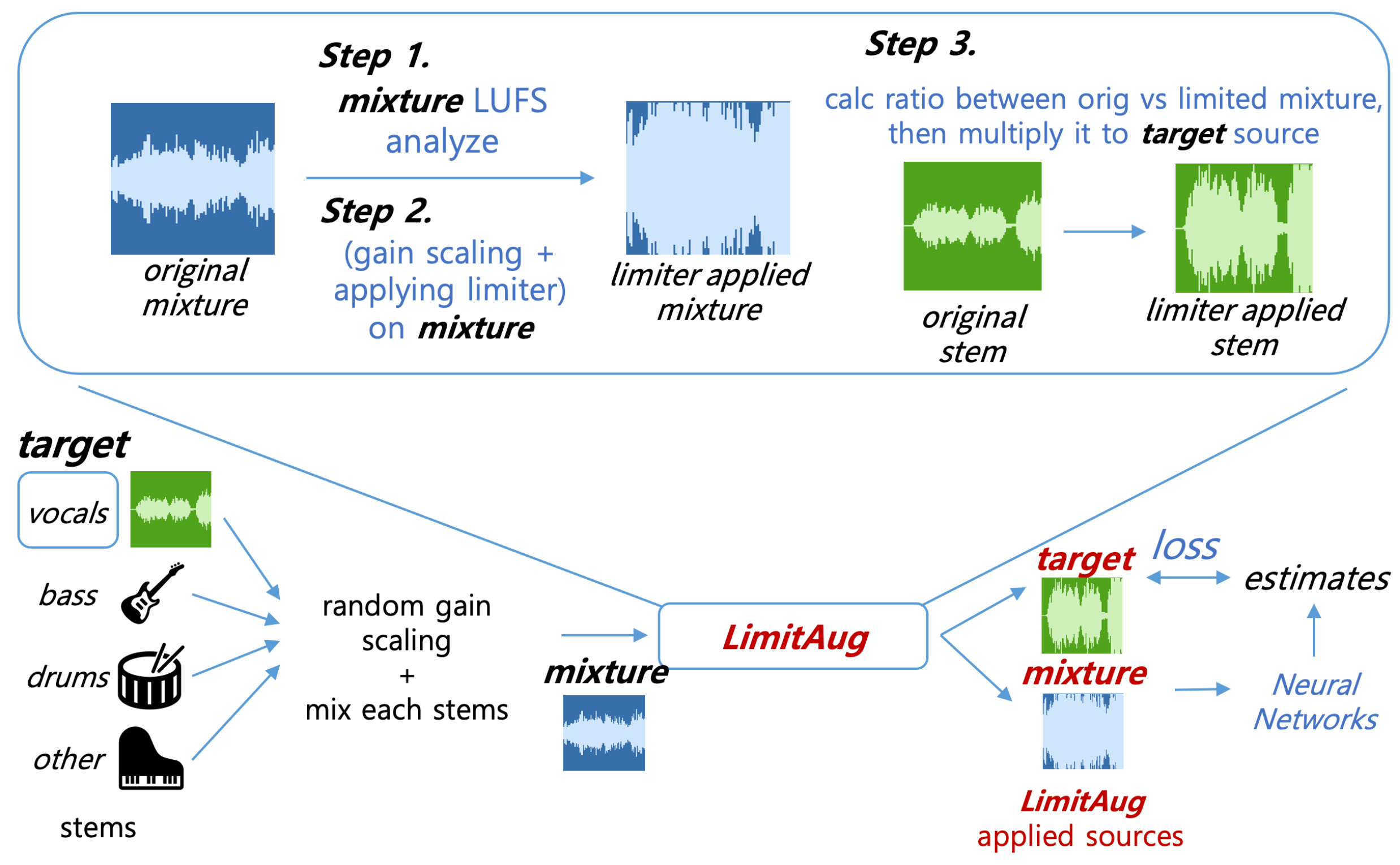
=> element-wise ratio calculation

=> Same technique applied on both musdb-XL and LimitAug



Methods

- real-world mastering finished music vs. standard train examples
  - => key differences
    - ***1)* overall amplitude scales**
    - ***2)* signal distortion caused by a limiter**
- How to avoid the domain mismatch?
  - => **Input loudness normalization (for *1)*)**
    - *i)* both in train and eval stage
    - *ii)* only at eval stage => if models are trained w/o loud-norm
  - => ***LimitAug* (for *2)*) and *LimitAug + loud-norm* (for both *1)* and *2)*)**
    - **- Let's use a limiter when making train examples**



Experiments

- Significant performance degradation of sota networks on *musdb-XL* datasets.

network	extra train data	test data	SDR median (mean) [dB]				
			vocals	*bass*	*drums*	*other*	*avg*
Open-unmix [6]	-	*hq*	6.16 (2.54)	5.03 (2.67)	6.00 (5.46)	4.22 (3.46)	5.35 (3.53)
		L	6.33 (1.63)	4.81 (2.71)	5.82 (5.38)	4.11 (3.42)	5.27 (3.28)
		XL	5.98 (0.89)	4.76 (2.59)	4.97 (4.89)	4.04 (3.29)	4.94 (2.92)
TFC-TDF -U-net [20]	-	*hq*	7.18 (4.26)	5.59 (3.35)	5.76 (5.30)	4.04 (3.18)	5.64 (4.02)
		L	7.03 (3.65)	5.41 (3.08)	5.52 (5.09)	3.67 (3.00)	5.41 (3.71)
		XL	6.95 (3.14)	5.48 (2.90)	5.11 (4.68)	3.55 (2.82)	5.27 (3.39)
Demucs v3-A [19]	-	*hq*	**8.11 (5.22)**	**9.34** (6.21)	**8.57 (8.01)**	**5.51 (5.03)**	**7.88 (6.12)**
		L	7.54 (5.15)	9.32 **(6.22)**	8.26 (7.65)	5.51 (5.01)	7.66 (6.01)
		XL	7.30 (4.86)	9.19 (6.14)	7.62 (6.78)	5.37 (4.97)	7.37 (5.69)
Open-unmix [6]	✓	*hq*	7.02 (4.93)	5.91 (4.06)	7.18 (6.91)	4.94 (4.76)	6.26 (5.17)
		L	6.83 (5.12)	6.23 (4.09)	7.07 (6.92)	4.94 (4.78)	6.27 (5.23)
		XL	6.70 (4.77)	6.16 (3.87)	6.80 (6.48)	4.89 (4.61)	6.14 (4.93)
Spleeter [21]	25000+	*hq*	6.51 (4.42)	4.77 (3.57)	6.00 (6.09)	4.22 (4.12)	5.38 (4.55)
		L	6.18 (3.90)	4.73 (3.34)	5.67 (5.94)	4.37 (4.03)	5.24 (4.30)
		XL	6.03 (3.38)	4.80 (3.13)	5.55 (5.52)	4.24 (3.91)	5.15 (3.98)
Demucs v3-B [19]	200+ including *musdb-hq* test set	*hq*	**9.24 (7.05)**	**11.65 (9.58)**	**11.73 (11.34)**	**7.83 (8.03)**	**10.11 (9.00)**
		L	9.05 (6.91)	11.61 (9.55)	11.05 (10.27)	7.83 (7.91)	9.88 (8.66)
		XL	8.76 (6.41)	11.56 (9.29)	9.22 (8.78)	7.52 (7.51)	9.26 (8.00)

- Simple loud-norm only at eval stage can greatly reduce performance decrease

network	extra train data	test data	SDR median (mean) [dB]			
			vocals	*bass*	*drums*	*other*
Demucs v3-A [19]	-	*hq*	**8.11** (5.22)	**9.34 (6.21)**	**8.57 (8.01)**	5.51 **(5.03)**
		L	8.05 **(5.23)**	9.25 (6.20)	8.47 (7.92)	**5.53** (5.02)
		XL	7.93 (5.03)	9.27 (5.92)	7.74 (7.44)	5.55 (4.91)
Demucs v3-B [19]	200+ including *musdb-hq* test set	*hq*	**9.24 (7.05)**	**11.65 (9.58)**	**11.73 (11.34)**	**7.83 (8.03)**
		L	9.19 (7.04)	11.64 (9.55)	11.68 (11.21)	7.82 (8.02)
		XL	9.13 (6.90)	11.56 (9.33)	11.32 (10.75)	7.74 (7.95)

network	extra train data	SDR median [dB]		
		hq	*L*	*XL*
Open-unmix [6]	-	5.35	5.32	5.25
TFC-TDF-U-Net [20]	-	5.64	5.62	5.51
Demucs v3-A [19]	-	**7.88**	**7.82**	**7.62**
Open-unmix [6]	✓	6.26	6.25	6.18
Spleeter [21]	✓	5.38	5.33	5.21
Demucs v3-B [19]	✓	**10.11**	**10.08**	**9.94**

- Loud-norm in both train and eval stages is helpful not only for *musdb-XL* but for standard *musdb-hq*
- *LimitAug + loud-norm* is also helpful

network	methods	linear gain increase	*LimitAug*	input loud-norm	target LUFS	SDR median (mean) [dB]			
						hq	*L*	*XL*	*avg*
TFC-TDF -U-Net [20]	baseline	-	-	-	-	5.64 (4.02)	5.41 (3.71)	5.27 (3.39)	5.44 (3.71)
	(1)	✓	-	-	$\mathcal{N}(\mu_L,\sigma_L^2)$	**5.90** (4.31)	5.86 (4.33)	5.73 (4.15)	5.83 (4.26)
					$\mathcal{N}(\mu_{XL},\sigma_{XL}^2)$	5.32 (3.43)	5.36 (3.62)	5.28 (3.49)	5.32 (3.51)
	(2)	-	✓	-	$\mathcal{N}(\mu_L,\sigma_L^2)$	5.79 (4.30)	**5.90 (4.41)**	5.74 (4.25)	5.81 (4.32)
					$\mathcal{N}(\mu_{XL},\sigma_{XL}^2)$	5.69 (3.93)	5.72 (4.22)	5.57 (4.15)	5.66 (4.10)
	(3)	-	-	✓	-14	5.89 **(4.38)**	5.87 (4.35)	**5.82 (4.25)**	**5.86 (4.33)**
	(4)	-	✓	✓	$\mathcal{N}(\mu_L,\sigma_L^2)$, -14	5.87 (4.25)	5.85 (4.21)	5.76 (4.16)	5.83 (4.21)
					$\mathcal{N}(\mu_{XL},\sigma_{XL}^2)$, -14	5.78 (4.27)	5.78 (4.26)	5.73 (4.20)	5.76 (4.24)

- *LimitAug + loud-norm* is especially better than *vocals* and *other* stems

network	methods	target LUFS	test data	SDR median (mean) [dB]				
				vocals	*bass*	*drums*	*other*	*avg*
TFC-TDF -U-Net [20]	baseline	-	*hq*	7.18 (4.26)	5.59 (3.35)	5.76 (5.30)	4.04 (3.18)	5.64 (4.02)
			L	7.03 (3.65)	5.41 (3.08)	5.52 (5.09)	3.67 (3.00)	5.41 (3.71)
			XL	6.95 (3.14)	5.48 (2.90)	5.11 (4.68)	3.55 (2.82)	5.27 (3.39)
	(3) loud-norm	-14	*hq*	7.35 **(4.76)**	**5.93** (3.61)	**5.91 (5.37)**	4.39 (3.79)	**5.89 (4.38)**
			L	7.32 (4.72)	5.91 (3.61)	5.85 (5.29)	4.39 (3.78)	5.87 (4.35)
			XL	7.26 (4.64)	5.91 **(3.62)**	5.68 (4.99)	4.42 (3.78)	5.82 (4.25)
	(4) LimitAug, loud-norm	$\mathcal{N}(\mu_L,\sigma_L^2)$, -14	*hq*	**7.59** (4.64)	5.75 (3.25)	5.63 (5.28)	4.50 (3.82)	5.87 (4.25)
			L	7.58 (4.61)	5.69 (3.21)	5.62 (5.22)	4.50 (3.82)	5.85 (4.21)
			XL	7.48 (4.55)	5.67 (3.29)	5.36 (4.99)	**4.51 (3.82)**	5.76 (4.16)

Conclusions

- Musdb-XL => Loud and compressed evaluation data, perhaps useful for industry?!
- *LimitAug* => Also useful for other researches such as automatic music mixing.