Paper Number: 3636

On loss functions and evaluation metrics for music source separation

Enric Gusó (MTG Universitat Pompeu Fabra)
Jordi Pons, Santiago Pascual & Joan Serrà (Dolby)

IEEE ICASSP 2022

ABSTRACT

We investigate which loss functions provide better separations via benchmarking an extensive set of those for music source separation. To that end, we first survey the most representative audio source separation losses we identified, to later consistently benchmark them in a controlled experimental setup.

We also explore using such losses as evaluation metrics, via cross-correlating them with the results of a subjective test. Based on the observation that the standard signal-to-distortion ratio metric can be misleading in some scenarios, we study alternative evaluation metrics based on the considered losses.

TIME-DOMAIN LOSSES

- $L1_{time}$, $L2_{time}$: standard regression losses
- $SISDR_{time}$: scale-invariant SDR
- $SDSDR_{time}$: scale-dependant SDR
- $LOGL1_{time}$, $LOGL2_{time}$: log-compressed variants of $L1_{time}$ and $L2_{time}$

DEEP-FEATURE LOSS

L2-based loss functions on embeddings of the separator estimations and the targets, using a fixed pre-trained audio model (in our case, VGGish trained on Audioset) .

ADVERSARIAL LOSS

We regularize $L2_{freq}$ loss with an unsupervised loss by training discriminators on a Least-Squares GAN setup, allowing us to use additional unpaired data.

SPECTROGRAM-BASED LOSSES

- $L1_{freq}$, $L2_{freq}$: standard regression losses on the magnitude estimations
- Phase-sensitive (PSA): targets scaled with `cos(mixture_phase - target_phase)`
- $L1_{mask}$, $L2_{mask}$: standard regression losses on the magnitude masks
- Dissimilarity: $L2_{freq}$-$sum_{targets}$`(L2(target, other_target))`
- $SISDR_{freq}$: reshape spectrograms as vectors and compute SISDR
- $LOGL1_{freq}$, $LOGL2_{freq}$: log-compressed variants of $L1_{freq}$ and $L2_{freq}$

EXPERIMENTAL SETUP

- We train Open-Unmix models on MUSDB18 dataset
- We evaluate:
  - Source to Distortion Ratio (SDR)
  - Mean Opinion Scores (MOS) of the most promising losses
  - Loss-based metrics (training objectives used as evaluation metrics) for cross-correlating with MOS

SDR RESULTS

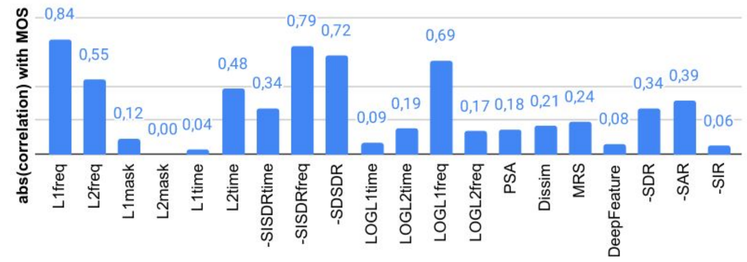
	Vocals	Drums	Bass	Other	Mean
$L1_{freq}$	5.95	5.58	4.24	3.90	4.92
$L2_{freq}$	6.40	5.86	5.28	**4.57**	**5.53**
$L1_{mask}$	5.10	4.14	2.88	2.53	**3.66**
$L2_{mask}$	4.74	4.39	2.41	2.71	**3.56**
$L1_{time}$	4.63	4.80	3.53	3.01	**3.99**
$L2_{time}$	4.55	4.25	3.06	2.92	**3.70**
$SISDR_{time}$	6.24	5.76	5.06	4.37	5.35
$SISDR_{freq}$	6.26	6.09	**5.55**	4.54	**5.61**
$SDSDR_{time}$	6.02	5.84	4.99	4.37	5.30

	Vocals	Drums	Bass	Other	Mean
$LOGL1_{time}$	5.95	5.82	5.23	4.35	5.34
$LOGL2_{time}$	5.88	5.81	5.04	4.10	5.21
$LOGL1_{freq}$	6.28	5.90	5.49	4.44	**5.53**
$LOGL2_{freq}$	6.15	5.79	5.38	4.36	5.42
PSA	6.18	**6.17**	5.10	4.33	5.44
$Dissim_{freq}$	6.04	5.66	5.17	4.38	5.31
MRS	5.82	5.11	4.48	3.57	4.80
DeepFeature	6.14	5.89	4.80	4.30	5.28
Adversarial	**6.50**	6.15	5.20	4.47	**5.58**

MEAN OPINION SCORES

	$L2_{freq}$	$SISDR_{freq}$	$LOGL1_{time}$	$LOGL1_{freq}$	Adv
Vocals	$57.24_{\pm 17.9}$	$53.33_{\pm 17.0}$	$54.83_{\pm 19.1}$	$\mathbf{59.10}_{\pm 17.0}$	$54.42_{\pm 18.4}$
Drums	$\mathbf{60.58}_{\pm 15.6}$	$59.60_{\pm 18.5}$	$59.10_{\pm 19.8}$	$59.76_{\pm 21.7}$	$54.40_{\pm 18.2}$
Bass	$61.62_{\pm 20.3}$	$62.80_{\pm 19.9}$	$57.19_{\pm 23.2}$	$62.13_{\pm 22.5}$	$\mathbf{64.28}_{\pm 21.0}$
Other	$\mathbf{56.90}_{\pm 18.6}$	$52.22_{\pm 19.0}$	$48.98_{\pm 18.5}$	$54.88_{\pm 18.5}$	$48.37_{\pm 20.4}$
Mean	$\mathbf{59.09}_{\pm 18.2}$	$56.99_{\pm 19.0}$	$55.01_{\pm 20.5}$	$58.96_{\pm 20.1}$	$55.39_{\pm 20.2}$

CORRELATION OF LOSS-BASED EVALUATION METRICS WITH MOS



DISCUSSION

- Objectively, best losses are: $L2_{freq}$, $SISDR_{freq}$, $LOGL1_{freq}$ and Adversarial.
- Subjectively, best losses are: $LOGL1_{freq}$ and $L2_{freq}$.
- Overall, we recommend $L2_{freq}$, $SISDR_{freq}$, $LOGL2_{freq}$ or $LOGL1_{freq}$.
- All losses lie in the upper-fair range of MOS. Still room for improvement.
- $L1_{freq}$ and $SISDR_{freq}$ correlate better with human judgement than SDR. It would be informative if future works also reported spectral distortion metrics like L1freq together with SDR.