# Key Metrics in Audio Source Separation

**Wonjun Park**
Arlington Computational Linguistics Lab
University of Texas at Arlington
Arlington, TX, USA
*wxp7177@mavs.uta.edu*

***Abstract*—Sound source separation is inevitable in the field of Animal Language Processing in terms of analyzing a target animal sound from a sound mixture collected from a record in noisy natural environments. An evaluation is a crucial step in determining whether the performance meets the requirements for conducting the specific task. In this paper, several metrics for sound separation are reviewed, expecting the enhancement of the knowledge about the metrics and further own ability to decide which metrics are suitable for the task, especially animal vocal separation.**

***Index Terms*—Animal language processing, audio source separation, evaluation metrics.**

## I. Introduction

Languages have been considered as a unique trait of humans, further sometimes the evidence of the intelligence. In the early stage of the languages, it is presumed that their primitives were started from a mean of alerting dangers to other humans in their groups, if predators came, for instance. This simple method to deliver information to others became so advanced that oral transmissions and records have been laid the foundation of the human intelligence.

With those human-centric perspectives on languages, recent books and studies [1]–[3] have started to move a regard into animals, breaking a new ground Animal Language Processing (ALP). Since no structural languages from animals are found, these studies mostly relied on the vocalization of the animals, somewhat contextual situations as well.

Noting that an assumption that designed environments like a laboratory at indoor can not draw out whole languages of animals is existed, however, the collection of clean audios which only a target animal is recorded is extremly hard, leading to the need of the animal vocal separation.

Especially, for example, while data collection approaches using Sound Event Detection (SED) are able to guarantee only a target vocalization is emerged on a range of an audio, a few overlapped vocalizations should be cut and abandoned, even if they were continuous sounds from the target.

A task, audio source separation, is consequently important in the field of ALP. In this paper, objective measurements [4] related to the task will be introduced;

- Signal-to-Distortion Ratio (SDR)
- Scale Invariant Signal-to-Distortion Ratio (SI-SDR)

## II. Metrics

### A. SDR

The definition of SDR has been changed over the years [5], after it firstly proposed in [6]. The current widely used formula [7], same as Signal-to-Noise Ratio (SNR), is defined as:

$$\text{SDR} = 10 \cdot \log_{10} \left( \frac{\|s\|^2}{\|s - \hat{s}\|^2} \right) \tag{1}$$

where $s$ is the target signal and $\hat{s}$ is the estimated signal.

Although the SDR is still used and reported, Le Roux et al., [4] pointed out that the metric is sensitive to the scale of the $\hat{s}$ which might not be the desired behavior in the Sound Source Separation task.

### B. SI-SDR

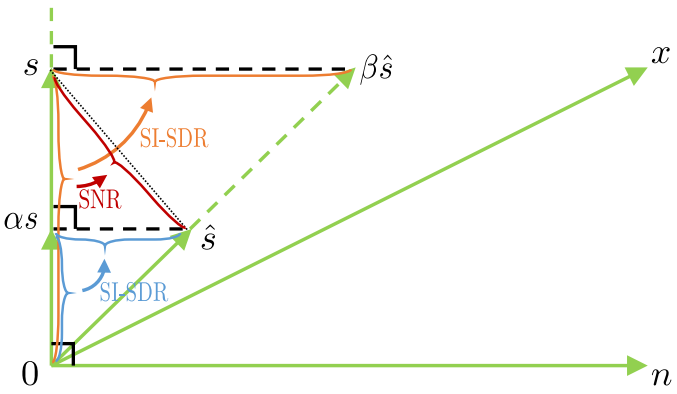


Fig. 1. Illustration of SI-SDR [4]

SI-SDR is proposed to overcome the limitation of SDR. A mixture $x$ is composed by $s$ and $n$, where $s$ is the target signal and $n$ is the noise. In a vector space, $s$ and $n$ construct their own orthogonal spaces. With the following formula, the SI-SDR is able to measure the robust metric without the scale variance.

$$\text{SI-SDR} = 10 \cdot \log_{10} \left( \frac{\|s\|^2}{\|s - \beta\hat{s}\|^2} \right) \tag{2}$$

$$\text{for } \beta = s \perp s - \beta\hat{s}$$

$$= 10 \cdot \log_{10} \left( \frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2} \right) \tag{3}$$

$$\text{for } \alpha = \text{argmin}_\alpha \|\alpha s - \hat{s}\|^2$$

### C. Other Metrics

While other metrics like Signal-to-Artifacts Ratio (SAR) and Signal-to-Interference Ratio (SIR) were also proposed in [6], since they are able to be derived from SDR, those are not widely used.

## REFERENCES


[1] E. Meijer, *Animal Languages: Revealing the Secret Conversations of the Living World*. John Murray, 2019.

[2] J. Huang, C. Zhang, M. Wu, and K. Zhu, "Transcribing vocal communications of domestic shiba lnu dogs," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13 819–13 832.

[3] T. Wang, X. Li, C. Zhang, M. Wu, and K. Zhu, "Phonetic and lexical discovery of canine vocalization," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 13 972–13 983.

[4] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[5] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, A. Défossez, M. Kim, W. Choi, C.-Y. Yu, and K.-W. Cheuk, "Music demixing challenge 2021," *Frontiers in Signal Processing*, vol. 1, p. 808395, 2022.

[6] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[7] F.-R. Stöter and A. Liutkus, "bsseval," https://github.com/sigsep/bsseval.