Local Equivariance Error-Based Metrics for Evaluating Sampling-Frequency-Independent Property of Neural Network

Kanami Imamura[1,2], Tomohiko Nakamura[2], Norihiro Takamune[1], Kohei Yatabe[3], and Hiroshi Saruwatari[1]
[1] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2] The National Institute of Advanced Industrial Science and Technology (AIST), Japan
[3] Department of Electrical Engineering and Computer Science,
Tokyo University of Agriculture and Technology, Tokyo, Japan

***Abstract*—Audio signal processing methods based on deep neural networks (DNNs) are typically trained only at a single sampling frequency (SF) and therefore require signal resampling to handle untrained SFs. However, recent studies have shown that signal resampling can degrade performance with untrained SFs. This problem has been overlooked because most studies evaluate only the performance at trained SFs. In this paper, to assess the robustness of DNNs to SF changes, which we refer to as the SF-independent (SFI) property, we propose three metrics to quantify the SFI property on the basis of local equivariance error (LEE). LEE measures the robustness of DNNs to input transformations. By using signal resampling as input transformation, we extend LEE to measure the robustness of audio source separation methods to signal resampling. The proposed metrics are constructed to quantify the SFI property in specific network components responsible for predicting time-frequency masks. Experiments on music source separation demonstrated a strong correlation between the proposed metrics and performance degradation at untrained SFs.**

***Index Terms*—Local equivariance error, sampling frequency, audio source separation, deep learning**

I. Introduction

Deep neural networks (DNNs) have been widely used for various audio signal processing tasks [1]. In these tasks, DNNs are typically trained and evaluated only at a single sampling frequency (SF). To apply the trained models to signals with different SFs, signal resampling is usually required. Although resampling is well-grounded in signal processing theory [2], recent studies have shown that it can degrade performance at untrained SFs in DNN-based audio source separation methods [3]–[7]. This observation underscores the need to assess not only separation performance at trained SFs but also how robust these methods are to SF changes—that is, how independent they are of SFs. We refer to this independence as *the SFI property* and address its evaluation in this paper.

Although there are well-established evaluation metrics for separation performance [8], [9], there is currently no standard metric for evaluating the SFI property. For separation performance, source-to-distortion ratio (SDR) [8] and scale-invariant SDR (SI-SDR) [9] are commonly utilized in research on audio source separation literature [10]. Previous studies have measured the SFI property by calculating these metrics at various SFs [3]–[7], [11], [12]. However, while this evaluation is useful for comparing the performance of audio source separation methods at each test SF, it does not quantify *how* SFI a method is. Since SDR and SI-SDR depend on SFs by definition, their values at different SFs cannot be directly compared. Therefore, we need to construct an evaluation metric that quantifies the SFI property and explains the performance degradation at untrained SFs.

In this paper, we propose evaluation metrics to quantify the SFI property by measuring the robustness of audio source separation methods to signal resampling. The proposed metrics are based on local equivariance error (LEE) [13]. Originally introduced for image processing tasks, LEE measures the robustness of DNNs to input transformations such as rotation and translation. Since signal resampling is also a form of input transformation, we extend LEE to quantify the SFI property. Through this extension, we find that directly applying LEE does not explain the performance degradation at untrained SFs, as discussed in Section IV-C. To resolve this issue, the proposed evaluation metrics target not the entire network but specifically the network components responsible for predicting time-frequency masks. Through experiments on music source separation, we will analyze the relationship between the proposed metrics and performance degradation at untrained SFs.

II. Related Works

A. Local Equivariance Error

LEE is defined on the basis of *equivariance*, a property of a function where a transformation applied to the input space results in an equivalent transformation in the output space. Equivariance in DNNs has proven effective for tasks such as image line drawing and semantic segmentation [14].

Let $f : \mathbb{V} \to \mathbb{V}'$ be a function and $x \in \mathbb{V}$ be its input sampled from a data distribution $\mathcal{D}$, where $\mathbb{V}$ and $\mathbb{V}'$ are input and output vector spaces, respectively. We say that f

This work was supported by JSPS KAKENHI under Grant JP23K28108 and JST ACT-X under Grant JPMJAX210G.

is equivariant under the action of a group $\mathcal{G}$ if the following equation holds for any $\mathrm{g} \in \mathcal{G}$:

$$f(\tau(\mathrm{g})x) = \tau'(\mathrm{g})(f(x)), \tag{1}$$

where $\tau(\mathrm{g}) : \mathbb{V} \to \mathbb{V}$ and $\tau'(\mathrm{g}) : \mathbb{V}' \to \mathbb{V}'$ are the representations of g for the input and output of f, respectively. For example, we consider the rotation group $\mathcal{G}$ for images and a group element g_r smoothly connected on $\mathcal{G}$ by the parameterization of angle r. The realizations of g_r in the input and output domains ($\tau(\mathrm{g}_r)$ and $\tau'(\mathrm{g}_r)$) are rotation matrices corresponding to angle r. When $r = 0$, they reduce to identity matrices, i.e., no rotation. For simplicity, we denote $\tau(\mathrm{g}_r)$ and $\tau'(\mathrm{g}_r)$ as g_r and g'_r.

LEE measures the degree of equivariance in a DNN. We assume that g'_r is invertible and define $h_r(x) = g'^{-1}_r(f(g_r(x)))$. The difference between $h_r(x)$ and $f(x)$ serves as a measure of deviation from equivariance. Taking the limit as $r \to 0$, LEE is given as

$$\mathrm{LEE}(f) = \frac{1}{V}\mathbb{E}[||\mathcal{L}f(x)||^2], \tag{2}$$

$$\mathcal{L}f = \lim_{r\to 0}\frac{h_r - f}{r} = \frac{d}{dr}h_r\bigg|_{r=0}, \tag{3}$$

where $\mathbb{E}$ denotes the expectation with respect to x over the distribution $\mathcal{D}$ and V is the output dimension of $f(x)$. $\mathcal{L}f$ is called the Lie derivative of f, which measures the degree of change in f under an infinitesimal transformation.

LEE was extended to approximately compute the contributions of specific layers using the chain rule of the Lie derivative [13]. We refer to this extension as *layerwise LEE*. Let $f_{\mathrm{NN}}(x) = (f_I \circ f_{I-1} \circ \ldots \circ f_1)(x)$ be a DNN composed of I layers, where f_i $(i = 1, \ldots, I)$ denotes the function representing layer i. By applying the chain rule, the Lie derivative of f_{NN} can be decomposed as

$$\mathcal{L}f_{\mathrm{NN}} = \sum_{i=1}^{I} J_{I:i+1}\mathcal{L}f_i, \tag{4}$$

where $J_{I:i+1}$ is the Jacobian matrix of $f_I \circ \ldots \circ f_{i+1}$ and $J_{I:I+1}$ is an identity matrix. For notational simplicity, we omit the input of each Lie derivative and Jacobian matrix. From (4), we derive the upper bound of the norm of $\mathcal{L}f_{\mathrm{NN}}$:

$$||\mathcal{L}f_{\mathrm{NN}}|| \leq \sum_{i=1}^{I} ||J_{I:i+1}\mathcal{L}f_i||. \tag{5}$$

Each term on the right-hand side can be interpreted as the contribution of an individual layer. The layerwise LEE of layer i is defined as the expectation of the term associated with f_i in (5) and serves as a tool for analyzing the influence of each layer on LEE. See [13] for further details on LEE and layerwise LEE.

B. TasNet Architecture

The TasNet architecture [15] is one of the foundational network architectures for DNN-based audio source separation and is widely utilized in literature [3], [16], [17]. It consists of an encoder f_{enc}, a decoder f_{dec}, and a mask predictor f_{mask}. The encoder and decoder serve as a trainable time-frequency transform and its inverse transform, respectively. The encoder transforms the input signal into a pseudo time-frequency representation, which is passed to the mask predictor. The mask predictor predicts masks of individual sources and multiplies the representation by these masks. The masked representations are then fed into the decoder to produce the separated signals. In this paper, we explore the DNN with the TasNet architecture and denote it as $f_{\mathrm{NN}} = f_{\mathrm{dec}} \circ f_{\mathrm{mask}} \circ f_{\mathrm{enc}}$.

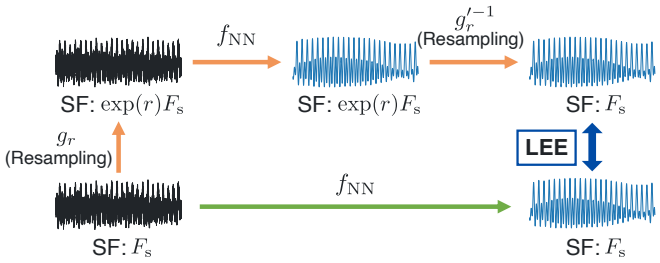


Fig. 1: Schematic of LEE using resampling as input transformation.

III. Proposed Metrics

A. LEE Using Resampling as Input Transformation

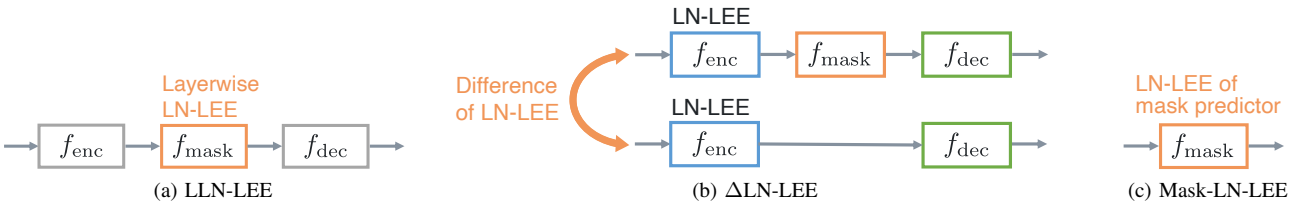
In this section, we first apply LEE to quantify the SFI property using resampling as the input transformation. Fig. 1 shows a schematic of LEE using resampling as the input transformation. For brevity, we reuse x, r, and g_r as the input signal, the parameter of SF change rate, and the signal resampling operation, respectively. The length and the SF of x are N and F_s, respectively. We use a windowed sinc interpolation for g_r to resample a signal sampled at F_s to $\exp(r)F_\mathrm{s}$. To compute the LEE value, the gradient of g_r with respect to r, denoted as $\partial g_r/\partial r$, must be computable. Given F_s, g_r can be represented as a linear transformation: $g_r(x) = Sx$. The (m, n)th entry of matrix $S \in \mathbb{R}^{\lceil \exp(r)N \rceil \times N}$ is given by

$$(S)_{m,n} = k\left(\frac{1}{F_\mathrm{s}}\left(\frac{m}{\exp(r)} - n\right), F_\mathrm{s}\right), \tag{6}$$

where $\lceil \cdot \rceil$ denotes the ceiling function and $k(t, F_\mathrm{s})$ is the following windowed sinc function:

$$k(t, F_\mathrm{s}) = z(t)\mathrm{sinc}(F_\mathrm{s}t), \quad \mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t}. \tag{7}$$

Here, $t \in \mathbb{R}$ denotes the continuous time and $z(t)$ is a window function. For $z(t)$, we choose a non-negative smooth window function and assume that $z(t) = 0$ for $t < -L/(2F_\mathrm{s})$ or $t > L/(2F_\mathrm{s})$, where L is a positive integer. From (6) and (7), if we choose $z(t)$ such that its gradient with respect to any t is computable, we can calculate $\partial g_r/\partial r$. That is, we can apply LEE to our problem. For example, we can use a Hann window for $z(t)$.



(a) LLN-LEE (b) ΔLN-LEE (c) Mask-LN-LEE

Fig. 2: Differences in DNN components evaluated in three proposed SFI evaluation metrics.

To match the characteristics of DNN-based audio source separation methods, we introduce two modifications to LEE for signal resampling. A preliminary experiment showed that the LEE value varied significantly with the scale of the DNN output. Hence, we first normalize the LEE value by the scale of the DNN output. Then, by taking the logarithm, similar to SDR, we define *log-normalized LEE (LN-LEE)* as

$$\text{LN-LEE}(f_{\text{NN}}) = \mathbb{E}\left[\log_{10}\frac{||\mathcal{L}f_{\text{NN}}(x)||}{||f_{\text{NN}}(x)||}\right]. \tag{8}$$

We hereafter treat LN-LEE as a baseline evaluation metric.

B. Motivation for Separate Evaluation of Mask Predictor

LN-LEE measures the robustness of DNNs to signal resampling. This robustness indicates that the performance of the DNN is consistent with SFs, i.e., the performance at untrained SFs does not degrade. However, contrary to this expectation, we observed a trend that LN-LEE decreases as the performance at untrained SFs decreases, as we will show in the experimental results of Section IV-C. This observation suggests that LN-LEE may not be suitable for evaluating the performance at untrained SFs.

To construct a metric that quantifies SFI properties and is consistent with the performance at untrained SFs, we focus on the influence of individual DNN components. As an example network architecture, we chose the TasNet architecture as mentioned in Section II-B. Since the encoder and decoder in this architecture correspond to time-frequency representations, the mask predictor is expected to have a greater impact on separation performance than the encoder and decoder. Thus, we focus on the mask predictor.

Motivated by this, we propose three evaluation metrics to quantify the SFI property of the mask predictor: *layerwise LN-LEE (LLN-LEE)*, *difference LN-LEE (ΔLN-LEE)*, and *LN-LEE of the mask predictor (Mask-LN-LEE)*. These metrics calculate the contribution of the mask predictor in different ways, and we will compare their effectiveness in Section IV. Although the proposed metrics are applicable to network architectures based on the short-time Fourier transform (STFT) [18]–[20], We leave the evaluation for STFT-based methods as a future work.

C. LEE-based Metrics Focusing on Mask Predictor

LLN-LEE: The first evaluation metric is based on layerwise LEE (see Fig. 2(a)). As in Section II-A, the Lie derivative of f_{NN} can be decomposed as

$$\mathcal{L}f_{\text{NN}} = \mathcal{L}f_{\text{dec}} + J_{\text{dec}}\mathcal{L}f_{\text{mask}} + J_{\text{dec}}J_{\text{mask}}\mathcal{L}f_{\text{enc}}. \tag{9}$$

Since the contribution of the mask predictor is the second term on the right-hand side of (9), the proposed metric is calculated similarly to LN-LEE:

$$\text{LLN-LEE}(f_{\text{NN}}) = \mathbb{E}\left[\log_{10}\frac{||J_{\text{dec}}(\mathcal{L}f_{\text{mask}})(f_{\text{enc}}(x))||}{||(f_{\text{mask}}\circ f_{\text{enc}})(x)||}\right]. \tag{10}$$

ΔLN-LEE: The second evaluation metric is based on the difference between the LN-LEE of the entire network and the network without the mask predictor (see Fig. 2(b)). In (9), we treat the term of the mask predictor and the other modules separately. From the norms of the term for the mask predictor and the other modules, we derive the following inequality:

$$||\mathcal{L}f_{\text{NN}}|| - ||\mathcal{L}f_{\text{dec}} + J_{\text{dec}}J_{\text{mask}}\mathcal{L}f_{\text{enc}}|| \le ||J_{\text{dec}}\mathcal{L}f_{\text{mask}}||. \tag{11}$$

The left-hand side serves as a lower bound for $||J_{\text{dec}}\mathcal{L}f_{\text{mask}}||$ and can be used to quantify the SFI property of the mask predictor. The left-hand side of (11) can be approximated using a network without the mask predictor, defined as $f_{\text{no\_mask}} = f_{\text{dec}} \circ f_{\text{enc}}$. The terms $\mathcal{L}f_{\text{dec}}$ and $J_{\text{dec}}J_{\text{mask}}\mathcal{L}f_{\text{enc}}$ correspond to the contributions of the encoder and decoder, respectively. Thus, $\mathcal{L}f_{\text{dec}} + J_{\text{dec}}J_{\text{mask}}\mathcal{L}f_{\text{enc}}$ can be approximated with the Lie derivative of $f_{\text{no\_mask}}$:

$$\mathcal{L}f_{\text{no\_mask}} = \mathcal{L}f_{\text{dec}} + J_{\text{dec}}\mathcal{L}f_{\text{enc}}. \tag{12}$$

With this approximation, the proposed metric is given as

$$\Delta\text{LN-LEE}(f_{\text{NN}}) = \text{LN-LEE}(f_{\text{NN}}) - \text{LN-LEE}(f_{\text{no\_mask}}). \tag{13}$$

Before taking the difference between $||\mathcal{L}f_{\text{NN}}(x)||$ and $||\mathcal{L}f_{\text{no\_mask}}(x)||$, we normalize the norm of each Lie derivative and take the logarithm (see (8)). An alternative approach is to first compute the difference of the norms and then normalize and take the logarithm of the result. However, preliminary experiments showed no significant difference between the two methods.

Mask-LN-LEE: The last evaluation metric is based on LN-LEE of the mask predictor (see Fig. 2(c)). We use the LN-LEE for f_{mask} as the proposed metric:

$$\text{mask-LN-LEE}(f_{\text{mask}}) = \mathbb{E}\left[\log_{10}\frac{||(\mathcal{L}f_{\text{mask}})(f_{\text{enc}}(x))||}{||(f_{\text{mask}}\circ f_{\text{enc}})(x)||}\right]. \tag{14}$$

Unlike LLN-LEE, this metric excludes the influence of f_{dec} because $(\mathcal{L}f_{\text{mask}})(f_{\text{enc}}(x))$ is not multiplied by J_{dec}.

IV. Experimental Analysis

A. Experimental Setup

We investigated the properties of the proposed evaluation metrics by applying them to music source separation methods. We used the same experimental setup for the music source separation methods as in [3]. We used the MUSDB18-HQ dataset [21], which consists of 150 tracks composed of vocals, bass, drums, and other. The SF of each track is 44.1 kHz. We used the official data split of 86, 14, and 50 tracks for training, validation, and test, respectively. The SF for the training and validation data was set to 32 kHz.

We used SFI Conv-TasNet [3] as a source separation model, which is an extension of the Conv-TasNet defined in [22] and is based on the TasNet architecture. It uses an SFI convolutional layer and an SFI transposed convolutional layer in the encoder and decoder, respectively. See [3] for details. For latent analog filters in the SFI convolutional layer and SFI transposed convolutional layer, we utilized modulated Gaussian filter (MGF). Its frequency response is given by

$$A(\omega) = \exp\left\{\frac{-(\omega-\mu)^2}{2\sigma^2} + \mathrm{j}\phi\right\} + \exp\left\{\frac{-(\omega+\mu)^2}{2\sigma^2} + \mathrm{j}\phi\right\}, \tag{15}$$

where ω is the angular frequency, j is the imaginary unit, μ is the center angular frequency, ϕ is the initial phase, and σ is the parameter corresponding to the bandwidth of the filter. We initialized μ and ϕ as in [3]. The initialization of σ is explained in Section IV-B. These parameters were trained jointly with other DNN parameters. We utilized the frequency-domain filter design in the SFI convolutional layer and the SFI transposed convolutional layer. The SFI Conv-TasNet was trained with a batch size of 12 for 250 epochs by RAdam [23] wrapped in a LookAhead optimizer [24]. We used the same data augmentations as in [3]. The loss function was the negative scale-invariant source-to-noise ratio (SI-SNR) [25].

For the evaluation of the separation performance, we used the SDR obtained with the BSSEval v4 toolkit [26]. Since the SDR is scale-dependent, we used the scaling method in [22]. To evaluate the impact of performance degradation due to signal resampling, we used the degradation from the SDR at the trained SF to the SDR at each SF. We call it the SDR degradation.

B. Setup of Proposed Evaluation Metrics

We compared four evaluation metrics for the SFI property: LN-LEE for the entire network (*entire LN-LEE*), LLN-LEE, ΔLN-LEE, and mask-LN-LEE. They were implemented on the basis of the official implementation[1] of LEE and calculated at the trained SF of 32 kHz. We extracted 5-second segments from each track in the test data. $z(t)$ was a Hann window and L was set to 24. To reduce the dependency on the initialization of the DNN, we used four random seeds to calculate each metric and used their average for comparison.

[1]https://github.com/ngruver/lie-deriv

To evaluate how well these metrics explain the performance degradation at untrained SFs, we used the correlation coefficients ρ between the metrics and the SDR degradation. The SDRs of SFI Conv-TasNet often decrease at SFs lower than the trained SF [3]. Hence, we calculated the SDR degradations at SFs of 8 and 16 kHz. In preliminary experiments, we observed that the SDR degradation increases as the initial value of σ in (15) becomes larger. Therefore, we varied the initial values of σ from 10π to 100π in increments of 10π.

C. Correlation of Proposed Metrics and Separation Performance

Figs. 3 and 4 show scatter plots of SDR degradation and each evaluation metric at 8 kHz and 16 kHz, respectively. Negative correlations were observed between the SDR degradation and entire LN-LEE for both SFs, demonstrating a contradiction with the desired trend. For the three proposed metrics that quantify the SFI property of the mask predictor, a positive correlation with the SDR degradation was observed across all metrics and SFs. These results demonstrate that the proposed metrics are more suitable for evaluating performance degradation at untrained SFs than entire LN-LEE. They also highlight the importance of focusing on the mask predictor to evaluate the performance degradation at untrained SFs.

Among the three proposed metrics, mask-LN-LEE achieved the highest correlation coefficients at 16 kHz. With a slight difference in correlation coefficient, ΔLN-LEE had similar values at 8 kHz and 16 kHz. From a computational perspective, mask-LN-LEE is more efficient than ΔLN-LEE because ΔLN-LEE requires an additional LN-LEE computation.

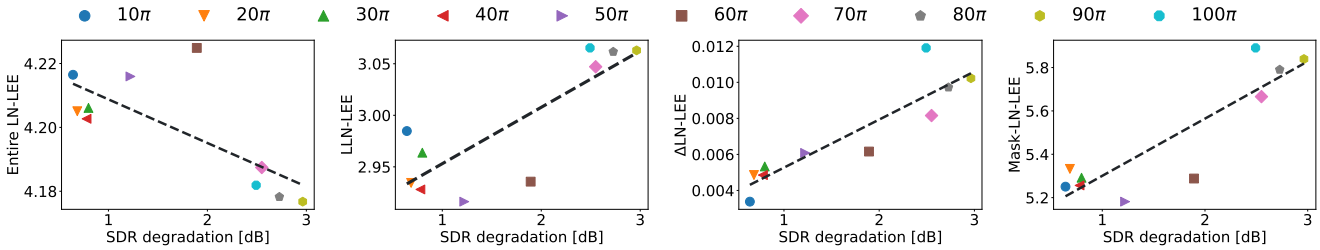
Importantly, the proposed metrics were computed not at the test SFs but at the trained SF. When we computed the proposed metric values at the test SFs, we found that the trends were similar to those in Figs. 3 and 4. This result indicates that the evaluation only at the trained SF is valid across various SFs.

V. Conclusion

To evaluate the SFI property of DNN-based audio source separation methods, we explored evaluation metrics based on LEE using resampling as an input transformation. Through this development, we found that directly applying LEE does not fully explain performance degradation at untrained SFs. To resolve this issue, we proposed three metrics to quantify the SFI property in the mask predictor in the TasNet architecture. Experiments on music source separation demonstrated that the proposed metrics are strongly correlated with the performance degradation at the untrained SFs. They also show the importance of focusing on the mask predictor to evaluate the SFI property. Extension of the proposed metrics to network architectures other than the TasNet architecture remains as future work.

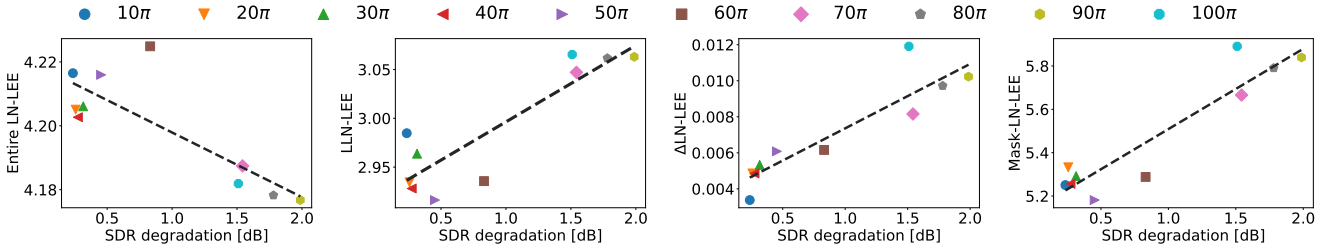
References

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Select. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019.



(a) Entire LN-LEE ($\rho = -0.74$) (b) LLN-LEE ($\rho = 0.83$) (c) ΔLN-LEE ($\rho = 0.91$) (d) Mask-LN-LEE ($\rho = 0.89$)

Fig. 3: Scatter plots of SDR degradation and each evaluation metric for SFI property at 8 kHz. Black dashed lines denote linear regression lines. Each point is average of metric values calculated using models trained with four random seeds.



(a) Entire LN-LEE ($\rho = -0.83$) (b) LLN-LEE ($\rho = 0.89$) (c) ΔLN-LEE ($\rho = 0.91$) (d) Mask-LN-LEE ($\rho = 0.93$)

Fig. 4: Scatter plots of SDR degradation and each evaluation metric for SFI property at 16 kHz. Black dashed lines denote linear regression lines. Each point is average of metric values calculated using models trained with four random seeds.

[2] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015.

[3] K. Saito, T. Nakamura, K. Yatabe, and H. Saruwatari, "Sampling-frequency-independent convolutional layer and its application to audio source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2928–2943, 2022.

[4] K. Imamura, T. Nakamura, N. Takamune, K. Yatabe, and H. Saruwatari, "Algorithms of sampling-frequency-independent layers for non-integer strides," in *Proc. Eur. Signal Process. Conf.*, 2023, pp. 326–330.

[5] J. Yu and Y. Luo, "Efficient monaural speech enhancement with universal sample rate band-split RNN," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, 5 pages.

[6] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, "Toward universal speech enhancement for diverse input conditions," in *Proc. Autom. Speech Recognit. Underst. Workshop*, 2023, 6 pages.

[7] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt *et al.*, "URGENT challenge: Universality, robustness, and generalizability for speech enhancement," in *Proc. INTERSPEECH*, 2024, pp. 4868–4872.

[8] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[9] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 626–630.

[10] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, "30+ years of source separation research: Achievements and future challenges," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2025, 5 pages.

[11] J. Paulus and M. Torcoli, "Sampling frequency independent dialogue separation," in *Proc. Eur. Signal Process. Conf.*, 2022, pp. 160–164.

[12] K. Imamura, T. Nakamura, K. Yatabe, and H. Saruwatari, "Neural analog filter for sampling-frequency-independent convolutional layer," *APSIPA Trans. Signal Inf. Process.*, vol. 13, no. 1, 2024.

[13] N. Gruver, M. A. Finzi, M. Goldblum, and A. G. Wilson, "The Lie derivative for measuring learned equivariance," in *Proc. Int. Conf. Learn. Represent.*, 2023, 9 pages.

[14] D. Chen, M. Davies, M. J. Ehrhardt, C.-B. Schönlieb, F. Sherry, and J. Tachella, "Imaging with equivariant deep learning: From unrolled network design to fully unsupervised learning," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 134–147, 2023.

[15] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 696–700.

[16] ——, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[17] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 46–50.

[18] Y. Luo and J. Yu, "Music source separation with band-split RNN," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1893–1901, 2023.

[19] W.-T. Lu, J.-C. Wang, Q. Kong, and Y.-N. Hung, "Music source separation with band-split RoPE transformer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 481–485.

[20] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[21] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019.

[22] D. Samuel, A. Ganeshan, and J. Naradowsky, "Meta-learning extractors for music source separation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 816–820.

[23] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2020, 13 pages.

[24] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: k steps forward, 1 step back," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9597–9608.

[25] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. INTERSPEECH*, vol. 8, 2016, pp. 545–549.

[26] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Var. Anal. Signal Sep.*, 2018, pp. 293–305.