A Multichannel Extension of Language-Queried Audio Source Separation via Independent Vector Analysis

Yuki Nakamura, Taishi Nakashima, and Nobutaka Ono
Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo, Japan
nakamura-yuki2@ed.tmu.ac.jp, taishi@tmu.ac.jp, onono@tmu.ac.jp

***Abstract*—In this paper, we propose a multichannel extension of Language-Queried Audio Source Separation (LASS) using Independent Vector Analysis (IVA). LASS enables the separation of arbitrary sound sources from a mixture based on natural language descriptions; however, conventional models assume a single-channel input and do not utilize spatial information. As one approach to extending LASS to multichannel processing, we consider utilizing LASS-separated signals as source models in Auxiliary-function-based IVA (AuxIVA), leveraging spatial and language query information for multichannel source separation. Furthermore, we investigate two approaches for integrating the LASS-separated signals into the IVA framework: a weighted arithmetic mean and a weighted geometric mean for mixing the source model variances. We demonstrate that the weighted geometric mean achieves higher separation performance through simulation experiments than the arithmetic mean. Experimental results indicate that the proposed method successfully extends LASS to multichannel source separation.**

***Index Terms*—blind source separation, language-queried audio source separation, multi-channel signal processing**

I. Introduction

Sound source separation is a technique for separating sound mixtures containing multiple sound sources into individual sound sources. Various methods have been proposed that are specific to different types of sound sources, such as music separation, speech enhancement, and acoustic event separation. Blind Source Separation (BSS) [1], [2] assumes multi-channel inputs and uses the spatial information in the mixed signals to separate the sound sources. A popular method includes Auxiliary-function-based Independent Vector Analysis (AuxIVA) [3]. One of the AuxIVA variants assumes time-varying Gaussian distribution with constant variances among frequencies as the source models [4].

On the other hand, universal sound separation (USS), which aims to separate arbitrary sound sources from various types of sound sources in real-world recordings, has been studied extensively in recent years [5]–[7]. In particular, Language-queried Audio Source Separation (LASS) [8], [9] enables source separation based on natural language descriptions. Most existing LASS models focus on single-channel separation. While they achieve strong separation performance, they do not explicitly leverage spatial information in the same way as conventional BSS methods.

Meanwhile, Neural Beamformer [10] is an example of a method that integrates single-channel time-frequency mask estimation into multichannel processing by incorporating it into beamformer design. This demonstrates that single-channel separation techniques can effectively combine with multichannel spatial processing to improve separation performance. In a similar way, model-based IVA [11]–[13] was proposed as a framework incorporating single-channel source separation methods into multichannel BSS to enhance separation accuracy. Although many methods that apply DNNs to multichannel source separation with full-rank spatial covariance models and local Gaussian models [14]–[18], extending LASS to multichannel processing enables flexible source separation guided by natural language queries while also providing practical advantages. Specifically, it eliminates the need for additional training of a multichannel model by leveraging a pretrained single-channel LASS model.

Therefore, in this paper, as one of the multichannel extensions of LASS, we propose a method that utilizes LASS-separated signals as source models for model-based IVA. To achieve effective integration, we introduce a novel approach for mixing source model variances using a weighted arithmetic mean and a weighted geometric mean. Through simulation experiments assuming time-invariant sound propagation, we demonstrate that the weighted geometric mean achieves higher separation performance than the arithmetic mean. These results indicate that our method provides a promising approach for extending LASS to multichannel processing by appropriately incorporating IVA, demonstrating an effective strategy for leveraging both spatial and language query information in source separation.

II. Related Work

A. Language-queried audio source separation

LASS is one of the frameworks of USS, which is a task to separate arbitrary sound sources from a mixture of various types of sound sources. AudioSep [9], a representative method of LASS, was also selected as the baseline algorithm for Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 Task 9, a mainstream international competition in the field of sound environment recognition. LASS is characterized by its ability to query specific sound sources using intuitive natural language descriptions and to include auxiliary information such as the temporal relationship of the target sounds. The LASS model consists of a query

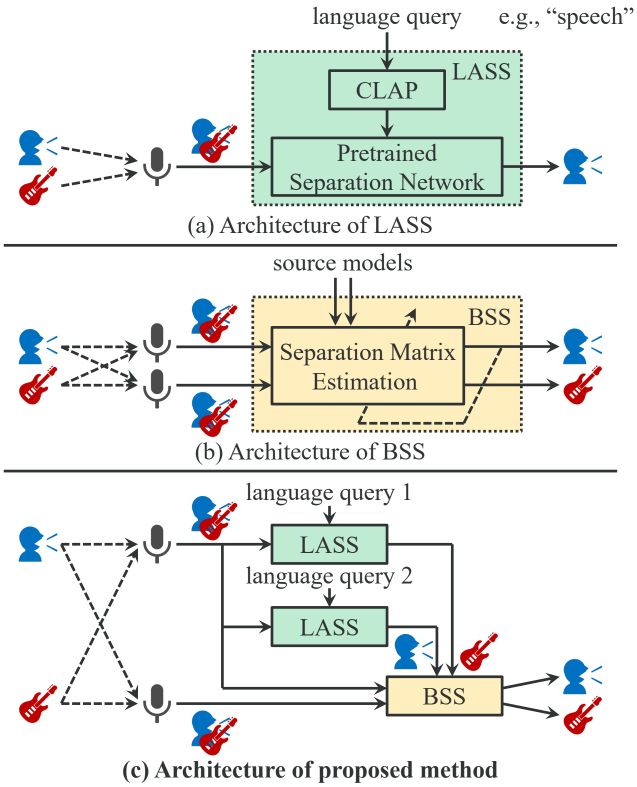


Fig. 1. Block diagram of conventional and proposed methods.

encoder and a separation network. The query encoder is a Contrastive Language-Audio Pre-training (CLAP) model [19] that has been pre-trained on labeled audio-text pair data. The separation network is trained to estimate the separation mask corresponding to the input text. A block diagram of LASS is shown in Fig. 1 (a).

B. Blind source separation

BSS is a task to estimate individual sound source signals from mixed signals. It does not require any information about the mixture, such as the positional relationship between the sound sources and microphones or spatial impulse response, and can be applied without pre-training. We assume a determined condition ($M \geq N$), where M is the number of microphones and N is the number of sources. A block diagram of the BSS is shown in Fig. 1 (b).

Let $\boldsymbol{x}(f,t) = [x_1(f,t), ..., x_M(f,t)]^\top \in \mathbb{C}^M$ be the mixed signals in the Short-Time Fourier Transform (STFT) domain, f be the discrete frequency bin index, and t be the discrete time frame index. Let $\boldsymbol{y}(f,t) = [y_1(f,t), ..., y_N(f,t)]^\top \in \mathbb{C}^N$ be the separated signals, and estimate the separation matrix $W(f) = [\boldsymbol{w}_1(f), ..., \boldsymbol{w}_N(f)]^{\mathsf{H}} \in \mathbb{C}^{M\times N}$ according to $\boldsymbol{y}(f,t) = W(f)\boldsymbol{x}(f,t)$ so that each element of the separated signals is statistically independent.

AuxIVA [3] is one of the popular BSS methods and alternately updates the weighted covariance matrix $V_i(f)$ and the separation matrix $W(f)$ as

$$V_i(f) = \frac{1}{T}\sum_{t=1}^{T} \frac{\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{\frac{1}{F}\sum_{f=1}^{F}|y_i(f,t)|^2}, \tag{1}$$

$$\boldsymbol{w}_i(f) \leftarrow (W(f)V_i(f))^{-1}\boldsymbol{e}_i, \tag{2}$$

$$\boldsymbol{w}_i(f) \leftarrow \frac{\boldsymbol{w}_i(f)}{\sqrt{\boldsymbol{w}_i^{\mathsf{H}}(f)V_i(f)\boldsymbol{w}_i(f)}}, \tag{3}$$

where i is the sound source index, T is the number of time frames, F is the number of frequency bins, $\cdot^{\mathsf{H}}$ is the complex conjugate transpose of the vectors, and $\boldsymbol{e}_i$ and $\boldsymbol{w}_i(f)$ are the i-th row vectors of the identity matrix and separation matrix, respectively.

Most conventional methods based on AuxIVA employ the time-varying Gaussian distribution as its source model [4] and assumes that the variance of the distribution is constant across frequencies. More flexible source model with different variance for each time-frequency component would improve separation performance. Therefore, model-based IVA [11], which uses the separated signals by single-channel source separation or binary masking as the source models, was proposed. In model-based IVA, the denominator in eq. (1) is replaced by the separated signal $z_i(f,t)$ as

$$V_i(f) = \frac{1}{T}\sum_{t=1}^{T} \frac{\boldsymbol{x}(f,t)\boldsymbol{x}^{\mathsf{H}}(f,t)}{|z_i(f,t)|^2}. \tag{4}$$

III. LASS MODEL-IVA

A. Procedure of LASS model-IVA

In this study, we propose a method of using the separated signals of LASS as source models for model-based IVA as one of the multi-channel extensions of LASS. A block diagram of the proposed method is shown in Fig. 1 (c). The procedure of the proposed method is as follows.

First, various types of sound sources are observed with multiple microphones. In this paper, we assume that all sources propagate from a single location to the microphones with time-invariant impulse responses. In other words, this is the same problem formulation assumed in the BSS.

Next, we apply LASS to the observed signals, each with a different language query. For example, if speech and music are observed, "speech" and "music" are given.

Finally, we apply IVA using the separated signal by LASS as the variances of the source models (LASS model-IVA) to the observed signals. Algorithm 1 summarizes the procedure of updating the separation and covariance matrices in LASS model-IVA. We explain the details of how we calculate the covariance matrix in the next subsection.

B. Mixing LASS variance and IVA variance

When calculating the variance of the IVA source model, if we simply give the separated signals by LASS as $z_i(f,t)$ in (4), the separation performance of the proposed method is directly affected by the one of LASS. Therefore, the final separation performance will be worse if the language query is not appropriate or if the LASS does not separate well.

Algorithm 1 Procedure of LASS model-IVA

```
Require: Set the number of iterations $N_\mathrm{i}$
Require: Set the mixing method of variance, arithmetic
    or geometric
Require: Initialize separation matrices $W(f)$ $(\forall f)$
$z_i(f,t) \leftarrow$ separated signal by LASS $(\forall i, f, t)$
for iter = 1 to $N_\mathrm{i}$ do
    $\boldsymbol{y}(f,t) = W(f)\boldsymbol{x}(f,t)$ $(\forall f, t)$
    for $i = 1$ to $N$ do
        $C_i^2(f) \leftarrow \frac{\sum_{t=1}^{T}|z_i(f,t)|^2}{\sum_{t=1}^{T}\frac{1}{F}\sum_{f=1}^{F}|y_i(f,t)|^2}$ $(\forall f)$
        if mixing method of variance is arithmetic then
            Caclculate $\sigma_i^2(f,t)$ $(\forall f, t)$ by (5)
        else if geometic then
            Caclculate $\sigma_i^2(f,t)$ $(\forall f, t)$ by (6)
        end if
        $V_i(f) \leftarrow \frac{1}{T}\sum_{t=1}^{T}\frac{\boldsymbol{x}(f,t)\boldsymbol{x}^\mathsf{H}(f,t)}{\sigma_i^2(f,t)}$ $(\forall f)$
        $\boldsymbol{w}_i(f) \leftarrow (W(f)V_i(f))^{-1}\boldsymbol{e}_i$ $(\forall f)$
        $\boldsymbol{w}_i(f) \leftarrow \frac{\boldsymbol{w}_i(f)}{\sqrt{\boldsymbol{w}_i^\mathsf{H}(f)V_i(f)\boldsymbol{w}_i(f)}}$ $(\forall f)$
    end for
    $\boldsymbol{y}(f,t) = W(f)\boldsymbol{x}(f,t)$ $(\forall f, t)$
end for
```

Thus, we propose an approach that uses a mixture of the variances obtained from the separated signals of the LASS and the variances used in conventional IVA as the variances of the source models. We expect more robust separations with this approach.

A conventional mixing method for variance is weighted arithmetic mean [20], [21]. However, we consider two types of variance for the different mixing methods: *weighted arithmetic mean*

$$\frac{1}{\sigma_i^2(f,t)} = \alpha\frac{1}{|z_i(f,t)|^2} + (1-\alpha)\frac{1}{C_i^2(f)\frac{1}{F}\sum_{f=1}^{F}|y_i(f,t)|^2}, \tag{5}$$

and *weighted geometric mean*

$$\frac{1}{\sigma_i^2(f,t)} = \frac{1}{(|z_i(f,t)|^2)^{\alpha}(C_i^2(f)\frac{1}{F}\sum_{f=1}^{F}|y_i(f,t)|^2)^{1-\alpha}}, \tag{6}$$

where α is a hyperparameter indicating the weight of the variance and $C_i^2(f)$ is a coefficient used to scale the separated signals by IVA and LASS calculated as

$$C_i^2(f) = \frac{\sum_{t=1}^{T}|z_i(f,t)|^2}{\sum_{t=1}^{T}\frac{1}{F}\sum_{f=1}^{F}|y_i(f,t)|^2}. \tag{7}$$

We then update the weighted covariance matrix as

$$V_i(f) = \frac{1}{T}\sum_{t=1}^{T}\frac{\boldsymbol{x}(f,t)\boldsymbol{x}^\mathsf{H}(f,t)}{\sigma_i^2(f,t)}. \tag{8}$$

Note that hereafter, LASS-model IVA also refers to a model using the variance mixing described in this subsection for simplicity.

IV. Experiments

A. Setup

Multi-channel mixed signals were created by simulation using Pyroomacoustics [22]. Note that we are assuming that all sound sources propagate from a single location to the microphones with time-invariant impulse responses. The number of sound sources and microphones were set to two each, with a distance of $2\,\mathrm{m}$ between sound sources and microphones, and a microphone distance of $0.02\,\mathrm{m}$. The directions of arrival of the sound sources were selected in pairs from a predefined set of angles: $30°$, $45°$, $60°$, $90°$, $120°$, $135°$, and $150°$. This results in 21 unique source direction pairs. The sampling frequency was $16\,\mathrm{kHz}$ and the reverberation time was $200\,\mathrm{ms}$. For STFT, a Hamming window of frame length 4096 was shifted with 1/2 overlap. We selected speech and music as sources for which the assumption that sound propagates to the microphone with a time-invariant impulse response from a single position is valid, and used 50 samples each from JSUT [23] and MUSDB18 [24], respectively. The LASS method was AudioSep-DP [25], a model that achieved first place in DCASE 2024 Task 9. In IV-B, IV-C, and IV-D the language queries were "female speech" and "musical instrument," respectively. An evaluation metric for the separated signals was the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [26], which was compared using the median value of all samples. We conducted the following four experiments to confirm the effectiveness of the proposed method.

B. Evaluating proposed method for each parameter

To select the optimal parameters for the proposed method, we compared the mixing method (arithmetic mean or geometric mean) and the mixing ratio ($\alpha = 0.1, 0.2, ..., 1.0$) of the variances with and without scaling by $C_i^2(f)$. The experimental results are shown in Table I. The results show that the proposed method achieves the highest SI-SDR improvement when the weight of variance of LASS is 0.4 and the mixing method is geometric mean, and there is no scaling by $C_i^2(f)$. This means that higher separation performance can be achieved by moderately mixing the variance of IVA updated by iterations, rather than using only the variance of LASS as the variance of the model. In addition, comparing the arithmetic mean and the geometric mean, the geometric mean achieved a higher overall improvement in SI-SDR. Comparing with and without scaling by $C_i^2(f)$, in the case of the arithmetic mean, scaling significantly improves the SI-SDR improvement. In subsequent experiments, we used the optimal parameters for the proposed method.

C. Evaluating model-IVA for each source model

To verify the efficacy of the separated signals by LASS as source models for IVA, we compared conventional model-based IVA methods with that of IVA using the separated signals by AudioSep-DP as source models (LASS model-IVA). Model-based IVA methods include time-varying Gaussian model (TG model-IVA) and Independent Low-Rank Matrix Analysis (ILRMA) [27]. Note that the well-known ILRMA is referred to as NMF model-IVA to maintain consistency with the names of other methods in this paper. To confirm the performance limits of model-IVA, we also compared IVA using the observed signals before mixing as the source models (hereafter referred to as Oracle model-IVA). Note that the oracle source signals are not available in real-world applications for Oracle model-IVA. The number of bases for NMF model-IVA was set to 5 and 10.

The separated signals of the four methods are compared in terms of SI-SDR improvement and plotted in a box-and-whisker diagram as shown in Fig. 2 (a). The results show that the proposed method, LASS model-IVA, achieves higher separation performance than TG model-IVA and NMF model-IVA. Therefore, the separated signals of AudioSep-DP are superior to the conventional model of AuxIVA as a source models.

D. Separation performance of LASS for each input signal

To further improve the separation performance, AudioSep-DP was applied to the mixed, TG model-IVA separated, and LASS model-IVA separated signals, respectively, and we compared the SI-SDR of the final separated signals. As shown in Fig. 2 (b), the SI-SDR of the separated signals by AudioSep-DP was lower when the separated signals by TG model-IVA was used as input than when the mixed signals was used as input. Even when the input was a separated signal by the LASS model-IVA, the applying of AudioSep-DP reduced the SI-SDR improvement of the separated signal. Thus, despite the improvement of SI-SDR from the mixed signal by the multi-channel BSS, the SI-SDR of the separated signal was inferior when AudioSep-DP was applied as post-processing. One possible cause of this is a mismatch between the training and evaluation data due to the fact that the training data for AudioSep-DP does not include the BSS separation signal.

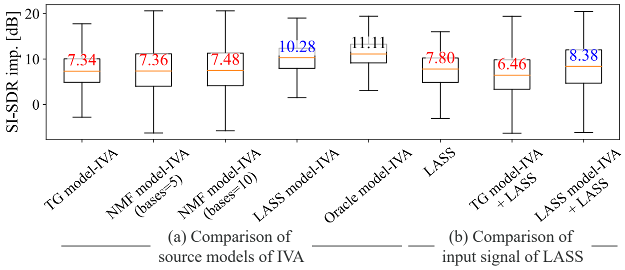
In addition, comparing the separation performance of the LASS model-IVA and LASS alone, the LASS model-IVA achieved higher SI-SDR improvement. Therefore, in the case of time-invariant propagation as assumed in this study, utilizing AudioSep-DP as a source models for model-IVA will improve the final separation performance, rather than simply applying it.

E. Separation performance of LASS for each caption of different lengths

To investigate the difference in the separation performance of LASS with caption length, we compared the separation performance of three methods using LASS (LASS, LASS model-IVA, and LASS model-IVA + LASS) with three types of captions generated using pre-trained Hierarchical Token-Semantic Audio Transformer (HTSAT) [28] as the language query. We calculated SI-SDRi independently for JSUT and MUSDB18. The average number of words in the captions of JSUT and MUSDB18 were [2.00, 5.00, 8.16], [1.88, 5.12, 7.54], in the order [short, middle, long], respectively. Example captions are shown in Table II.

TABLE I
RESULTS OF EXPERIMENTS IN IV-B. BOLD IS THE HIGHEST SI-SDR IMPROVEMENT [dB].

α	w/o $C_i^2(f)$		w/ $C_i^2(f)$	
	arithmetic	geometric	arithmetic	geometric
0.1	8.127	8.571	8.682	8.649
0.2	7.750	9.775	8.632	9.833
0.3	7.364	10.18	8.551	10.20
0.4	7.186	**10.28**	8.457	**10.27**
0.5	6.905	10.06	8.346	10.07
0.6	6.049	9.728	8.258	9.737
0.7	4.104	9.312	8.158	9.324
0.8	4.104	8.967	8.104	8.966
0.9	6.345	8.653	8.032	8.653
1.0	7.909	8.420	7.912	8.420



(a) Comparison of source models of IVA

(b) Comparison of input signal of LASS

Fig. 2. Results of experiments in IV-C and IV-D. The number in the box-and-whisker diagram is the median SI-SDR improvement [dB]. Red is the conventional method. Blue is the proposed method. TG: Time-varying Gaussian. NMF: Nonnegative Matrix Factorization. LASS: AudioSep-DP.

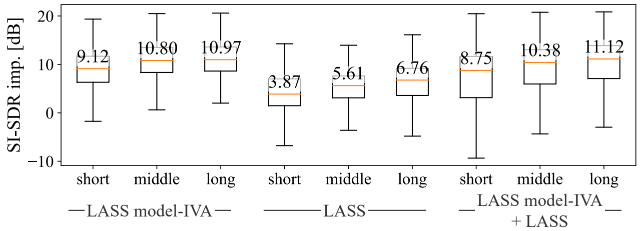


Fig. 3. Results of experiments in IV-E. The number in the box-and-whisker diagram is the median SI-SDR improvement [dB].

The experimental results are shown in Fig. 3. The results show that the SI-SDR improvement for all three methods improved as the number of words in the caption increased. This is thought to be due to the fact that MUSDB18 is composed of sound sources containing multiple instrumental sounds, and the increased number of words allows for a better representation of the sound sources.

V. Conclusion

In this paper, we proposed a multichannel extension of LASS using AuxIVA. While conventional LASS models focus on single-channel separation, our approach leverages LASS-separated signals as source models in IVA, incorporating both spatial and language query information for multichannel source separation.

TABLE II
EXAMPLE CAPTIONS FOR EACH OF THE THREE CAPTION LENGTH TYPES.

Dataset	Caption length	Example caption
JSUT [1]	short	a female
	middle	a female voice is speaking
	long	a female voice is saying a phrase
MUSDB18 [24]	short	music is
	middle	music is playing
	long	music is playing in a large room or hall

To effectively integrate LASS-separated signals into IVA, we introduced a mixing source model variances using a weighted arithmetic mean and a weighted geometric mean. Through simulation experiments, we demonstrated that the weighted geometric mean achieves higher separation performance than the arithmetic mean, providing a more effective strategy for variance estimation in model-based IVA.

These results indicate that our method successfully extends LASS to multichannel processing while ensuring efficient integration with IVA. By leveraging pretrained single-channel LASS models, our approach eliminates the need for additional multichannel model training and offers a computationally efficient alternative to deep learning-based multichannel separation methods.

ACKNOWLEDGEMENT

This work was supported by JST SICORP JPMJSC2306.

REFERENCES

[1] S. Makino, Ed., *Audio Source Separation*. Springer International Publishing, 2018.

[2] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Trans. SIP*, vol. 8, no. 1.

[3] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[4] ——, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*. IEEE, 2012, pp. 1–4.

[5] W. Rao, Y. Fu, Y. Hu, X. Xu, Y. Jv, J. Han, Z. Jiang, L. Xie, Y. Wang, S. Watanabe, Z.-H. Tan, H. Bu, T. Yu, and S. Shang, "Conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 679–686.

[6] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, "Toward universal speech enhancement for diverse input conditions," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.

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

[8] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. Interspeech*, 2022.

[9] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE/ACM Trans. ASLP*, pp. 1–15, 2024.

[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[11] A. R. López, N. Ono, U. Remes, K. Palomäki, and M. Kurimo, "Designing multichannel source separation based on single-channel source separation," in *Proc. ICASSP*, 2015, pp. 469–473.

[12] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. ICASSP*, 2021, pp. 176–180.

[13] H. Nammoku, K. Yamaoka, T. Nakashima, Y. Wakabayashi, and N. Ono, "Analysis and source separation of overlapping speech using corpus of everyday japanese conversation," in *Proc. ICA*, 2022.

[14] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1652–1664, 2016.

[15] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. ASLP*, vol. 27, no. 10, pp. 1601–1615, 2019.

[16] K. Saijo and R. Scheibler, "Independence-based joint dereverberation and separation with neural source model," in *Proc. Interspeech*, 2022, pp. 236–240.

[17] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural beamforming for moving speakers with self-attention-based tracking," *IEEE/ACM Trans. ASLP*, vol. 31, pp. 835–848, 2023.

[18] U.-H. Shin, S. Lee, T. Kim, and H.-M. Park, "Separate and reconstruct: Asymmetric encoder-decoder for speech separation," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=99y2EfLe3B

[19] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023, pp. 1–5.

[20] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proc. ICASSP*. IEEE, 2012, pp. 2417–2420.

[21] T. Hasumi, T. Nakamura, N. Takarnune, H. Saruwatari, D. Kitamura, Y. Takahashi, and K. Kondo, "Multichannel audio source separation with independent deeply learned matrix analysis using product of source models," in *Proc. APSIPA*. IEEE, 2021, pp. 1226–1233.

[22] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[23] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," 2017. [Online]. Available: https://arxiv.org/abs/1711.00354

[24] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[25] H. Yin, J. Bai, Y. Xiao, H. Wang, S. Zheng, Y. Chen, R. K. Das, C. Deng, and J. Chen, "Exploring text-queried sound event detection with audio source separation," in *Proc. ICASSP*, 2025, pp. 1–5.

[26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.

[27] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. ASLP*, vol. 24, no. 9, pp. 1626–1641, 2016.

[28] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. ICASSP*, 2022, pp. 646–650.