Multichannel Separation and Classification of Sound Events

Robin Scheibler, Tatsuya Komatsu, and Masahito Togami
LINE Corporation, Tokyo, Japan
Email: {robin.scheibler,komatsu.tatsuya,masahito.togami}@linecorp.com

***Abstract*—We investigate the use of determined blind source separation for sound event detection (SED) and classification using multichannel recordings. Our proposed system appends a single channel SED model to each of the output channels of the separation algorithm. We expect the number of events per channel to be reduced and overall performance increased. Such a system allows different number of channels at training and test time. We demonstrate the performance on the DCASE 2020 Sound Event Localization and Detection dataset. We compare baseline training on single channel recordings to using different combinations of 1, 2, 3, or 4 channel recordings. For the separation, we compare both a traditional source prior and a neural prior, trained without groundtruth signals available. First, we show that with the former, performance increases when using more channels. In addition, mixing different number of channels during training yields a system that is robust across varying number of channels. Second, while the training scheme proposed for the tailored source prior is not found effective for separation, it seems to be effective for data augmentation. This indicates that multichannel training data is beneficial, even when the target systems are single channel.**

I. Introduction

Enabling machines to listen to and identify environmental sounds has deep implications for accessibility [1], home monitoring [2], [3], and security [4], [5]. This task is generally known as sound event detection (SED) and is a lively area of research [6]. It is defined as labeling semantic events and marking their temporal location and duration in a signal. With increasing data set sizes, deep neural network (DNN) based multi-label classification models have received much attention, e.g., convolutional neural networks (CNNs) [7], [8] and long short-term memory (LSTM) [9], [10]. Convolutional recurrent neural networks (CRNNs) [11], [12] have become a strong baseline for neural approaches. More recently, self-attention based models including Transformer [13] and Conformer [14] have shown significant improvement in SED performance.

A particularity of SED is that it assumes that events can co-occur in time, i.e., it is a multi-label classification problem. One may conjecture that performing SED on the individual sounds co-occurring separately might be easier than dealing with their mixture. Hence, some approaches employ source separation techniques as a pre-processing step, for example, non-negative matrix factorization [15], [16]. This approach was promoted for Task 4 of the DCASE2020 challenge ("Sound event detection and separation in domestic environments"). However, only eight of the 54 proposed systems made use of source separation, with the best one among those ranking 15th. This is an indication that the role and benefits of separation in SED is not yet fully understood.

All the approaches to SED described so far operate on single channel recordings. However, the benefits of microphone arrays for sound processing have long been recognized [17]. They can sense the spatial cues of the impinging sound waves and have been widely applied for enhancement via beamforming [18], direction of arrival estimation (DOA) [19], and blind source separation (BSS) [20]. First steps towards using spatial cues in SED have been taken in Task 3 of the DCASE2019 and DCASE2020 Challenges, Sound Event Localization and Detection (SELD). The goal of this task is to perform jointly SED and DOA estimation of the detected events. It is the only task of the DCASE challenge using multichannel recordings [21]. The spatial information is contained in the phase and amplitude differences between the channels. Different methods have been proposed to integrate this information in neural architectures for SELD, for example, inter-channel time difference features [22], [23], histograms of DOA over the time-frequency plane [24], or a gated linear unit based network [25]. While all these prior works have demonstrated the potential of multichannel recordings to identify DOA of sound events, they have not explored their potential to improve the *detection performance itself*.

In this work, we propose to use multichannel BSS as a pre-processing step to improve SED performance. Independence-based BSS allows to perform a linear separation, introducing minimal distortion [26]. We expect that after separation, each output channel may contain only a single sound event that will be easier to identify. We set out to verify this hypothesis and measure the gain obtained. For the separation, we use independent vector analysis (IVA) which allows to deal with convolutive mixtures [27], [28] and is efficiently implemented via majorization-maximization [29], [30]. This proposed scheme based on BSS has the advantages to be agnostic to the microphone locations (which need not be known), and that the number of channels may be different at training and test time.

Our contributions are as follows. Using the DCASE2020 Task 3 dataset [21], we train a baseline single channel SED classifier. We also train classifiers on the same dataset, but where IVA is applied to the input data prior to classification. In the multichannel case, classification is done in parallel on all output channels with the same classifier, and the results aggregated by taking the maximum between all the channels.

For the separation, we investigate both a traditional, data agnostic, source model, and a DNN source model trained for the task at hand [31]. For the latter, we propose a training strategy that does not require groundtruth separated signals, but only the existing Task 3 dataset. We try several ways of combining the channels during training.

First, we find that separation is indeed effective to improve the performance. Training of 4 channels recordings to which IVA with the data agnostic prior has been applied yields the best performance on 4 channel test data. In this case, the F1 score is about $4\,\%$ more than the baseline single channel classifier. In addition, the same model can be applied to 2 and 3 channels recording with good performance. Our second finding concerns the trained source prior for IVA. In this case, we find that the training scheme proposed failed to yield an effective source model, most likely due to confounding noise present in the training signals. However, the SED classifier trained with it exhibits the best single channel performance overall. This indicates that remixing multichannel recordings may serve as an effective data augmentation scheme.

The rest of this paper is organized as follows. Section II introduces necessary elements of SED and BSS. We describe our proposed methodology in Section III. Experiments and their results come in Section IV, and we conclude in Section V.

II. Background

We start by introducing the necessary background for blind source separation with IVA. We denote vector and matrices by bold lower and upper case letters, respectively. Furthermore, $\boldsymbol{A}^\top$ and $\boldsymbol{A}^\mathsf{H}$ are the transpose and conjugate transpose, respectively, of matrix $\boldsymbol{A}$.

Consider a reverberant mixture of K sources recorded by M microphones. After applying the short-time Fourier transform (STFT), a good model for the multichannel signal is,

$$\boldsymbol{x}_{fn} = \sum_{k=0}^{K} \boldsymbol{A}_f \boldsymbol{s}_{fn}, \quad f = 1, \ldots, F,\ n = 1, \ldots, N, \tag{1}$$

where $\boldsymbol{x}_{fn} \in \mathbb{C}^M$ is the vector containing in its entries the components at frequency f and time index n of each microphones. The vector $\boldsymbol{s}_{fn}$ is similarly constructed but contains the components of each of the K sources. The entries of the *mixing matrix* $\boldsymbol{A}_f$ contain the transfer functions between sources and microphones at frequency f. For example, $(\boldsymbol{A}_f)_{mk}$ is the fth component of the discrete Fourier transform of the impulse response between source k and microphone m.

A. Independent Vector Analysis

Separation by IVA proceeds with finding a squared *demixing matrix* $\boldsymbol{W}_f \in \mathbb{C}^{M\times M}$, per frequency f, that when applied to the input signal produces outputs that are statistically independent. Let $\boldsymbol{w}_{fm}^\mathsf{H}$ be the mth row of $\boldsymbol{W}_f$. Then, we define the outputs of the separation as the M spectrograms,

$$\boldsymbol{Y}_m \in \mathbb{C}^{F\times N}, \quad \text{such that } y_{mfn} = (\boldsymbol{Y}_m)_{fn} = \boldsymbol{w}_{fm}^\mathsf{H} \boldsymbol{x}_{fn}. \tag{2}$$

Thus, IVA tries to ensure that the joint distribution of the output spectrograms is the product of their marginals, i.e.,

$$p_{1,\ldots,M}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M) = \prod\nolimits_{m=1}^{M} p(\boldsymbol{Y}_m), \tag{3}$$

where $p(\,.\,)$ is the probability density function of the outputs.

This strategy is concretely implemented by choosing a model for $p(\,.\,)$ and applying maximum likelihood estimation. This yields the following objective function,

$$\ell(\mathcal{W}) = \sum\nolimits_m G(\boldsymbol{Y}_m) - 2N \sum\nolimits_f \log|\det \boldsymbol{W}_f| \tag{4}$$

where $\mathcal{W} = \{\boldsymbol{W}_f\}_{f=1}^F$ and $G(\boldsymbol{Y}) = -\log p(\boldsymbol{Y})$. AuxIVA, [29], is an efficient algorithm to minimize (4) based on majorization-minimization [32]. It can be applied when $G(\boldsymbol{Y})$ admits a surrogate function $G^+(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$ such that

$$G(\boldsymbol{Y}) \le G^+(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum\nolimits_{fn} u_{fn}(\hat{\boldsymbol{Y}})|y_{fn}|^2 + c(\hat{\boldsymbol{Y}}), \tag{5}$$

with equality if and only $\boldsymbol{Y} = \hat{\boldsymbol{Y}}$, and where $u_{fn} : \mathbb{C}^{F\times N} \to \mathbb{R}_+$ and $c : \mathbb{C}^{F\times N} \to \mathbb{R}$. In this work, we use the *iterative source steering* variant of AuxIVA that applies a series of rank-1 update to the demixing matrix [30],

$$\boldsymbol{W}_f \leftarrow \boldsymbol{W}_f - \boldsymbol{v}_{mf}\boldsymbol{w}_{mf}^H, \tag{6}$$

for $m = 1, \ldots, M$, in order, with weight vector $\boldsymbol{v}_{mf}$,

$$(\boldsymbol{v}_{mf})_k = \begin{cases} \frac{\sum_n r_{kfn}\, y_{kfn}\, (y_{mfn})^*}{\sum_n r_{kfn}\, |y_{mfn}|^2} & \text{if } k \neq m \\ 1 - \left(\frac{1}{N}\sum_n r_{kfn}|y_{mfn}|^2\right)^{-1/2} & \text{if } k = m \end{cases} \tag{7}$$

where $r_{kfn} = u_{fn}(\boldsymbol{Y}_k)$. This iterative algorithm can be unrolled to obtain the structure described in Fig. 1 as part of the whole system diagram. In this structure, it is clear that contributions from the source, described by $u_{fn}(\boldsymbol{Y})$ and spatial model, i.e., (7), are cleanly separated. Several source priors are possible. For example, a time-varying Gauss prior gives $u_{fn}(\boldsymbol{Y}) = (\sum_{f'} |y_{f'n}|^2)^{-1}$ [33]. It is also possible to replace it with a DNN and learn its weights by backpropagating a separation loss through the iterations [31]. This last strategy requires access to the clean separated signals. Although these are not provided by the DCASE2020 Task 3 dataset, an alternative scheme is described in Section III-C.

III. Proposed Methodology

SED is overwhelmingly trained with and applied to single channel audio. In this paper, we propose to use IVA to handle the multichannel part. This strategy allows us to use a single channel SED model with multichannel data. Fig. 1 illustrates the overall structure of the proposed system, which can be summarized as cascading IVA and SED, combined to some tailored training strategies outlined in this section. Fig. 1 also shows the details of the different parts of the system. The separation operation is expected to reduce the number of events present in a single channel at the output. We then apply SED to each output channel separately. Finally, the output probabilities for each channels are combined as,

$$\boldsymbol{p}_{e,n} = \max_{m=1,\ldots,M} \boldsymbol{p}_{e,mn}, \tag{8}$$

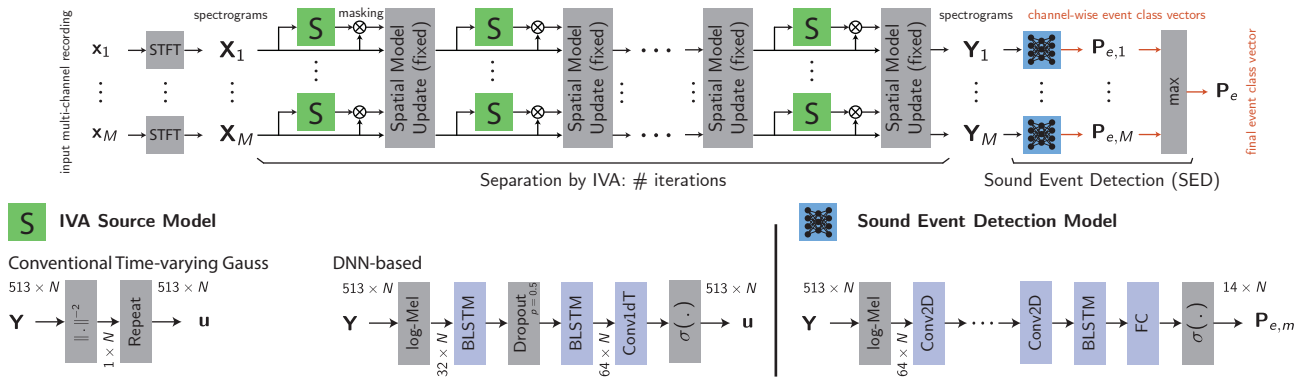

Fig. 1. Structure of the proposed separation/classification network. On top, the overall system including both IVA and SED. On the bottom, details of the architectures of the separation source model, and SED model. Blue blocks contain trainable parameters.

where $\boldsymbol{p}_{e,mn} \in [0,1]^C$ is the event probability vector for channel m and time index n, and C is the number of classes. The rational to use the $\max$ is that if separation is successful, an event might be detected in only one channel. If an event is not present, then all channels should return a low value for the probability, which is not increased by the $\max$ operation.

The final SED model that we obtain can be used with any number of channels. This allows several use case that would not be possible with a network architecture integrating both separation and SED. Prominently, we can train with multi-channel data and deploy on devices with a single microphone.

A. Architecture of the SED Model

The proposed method employs a CRNN-based SED model, which takes a sequence of log Mel-spectrogram features as input, and outputs the event probabilities $\mathbf{P}_e = [\mathbf{p}_{e,1}, ..., \mathbf{p}_{e,n}, ..., \mathbf{p}_{e,N}]$, $\mathbf{p}_{e,n} \in [0,1]^C$. Here, N and C are the number of time frames and event classes, respectively. The SED block consists of multiple CNN blocks, a BLSTM layer and a fully connected layer with sigmoid activation. First, input audio clips are transformed into log-mel spectrograms and fed to the CNN blocks. The BLSTM layer transforms the outputs of the CNN blocks by considering temporal dependencies and obtains event features. The fully connected layers with sigmoid activation acts as an event classifier and estimate the event probability. The event feature of each frame is input to the event classifier, and the event probabilities of each frame $\mathbf{p}_{e,n}$ are obtained. The model architecture is illustrated in Fig. 1. In the proposed scheme, this structure is repeated on each of the output channels of the separation, with shared weights.

B. Training Strategies

a) Baseline Approach: The most straightforward way to combine SED and IVA is to train on single channel data. At evaluation time, we pre-process the data with IVA and apply the single channel SED model on each output channel separately. This way, we expect that each output channel will contain the same number of events, or less, which should be beneficial. This approach only requires single channel training data, and is applicable to both single and multi-channel data at evaluation time. Its main benefit is to be able to re-use any existing model for SED with multi-channel data. One drawback is that the separation method may introduce some artefacts not present in the original training data and decrease the overall accuracy, instead of improving it.

b) Learning with Separated Events: The second strategy we propose is to apply separation during training as well. By doing so, the SED model will learn any peculiarities of the separation method, such as how to handle permutations often occurring with IVA [20]. We first run a fixed number of IVA iterations and then apply SED to each channels separately, as explained at the beginning of this section. With multichannel data, we also have the opportunity to train with different number of channels. With M-channels training data available, we can also train on all the distinct subsets of $M' < M$ channels. Furthermore, training can be done only for a fixed number of channels, or by including several number of channels. This allows to train models that can be used on devices with different number of channels.

C. Training a Separation Model

IVA has traditionally used a data independent source model such as time-varying Gauss [33], or Laplace [29]. Recently, it has been proposed to replace $u_{fn}(\boldsymbol{Y})$ in (5) by a DNN [31]. Its weights are then trained by back-propagating the scale-invariant signal-to-distortion ratio loss. This strategy could allow to create a separation model tailored specifically for the sound events that should be detected. However, it requires the groundtruth separated events, which are not available for the DCASE2020 SELD dataset [21]. Instead, we use the annotations of the dataset to pick segments where only one event occurs and remix them to obtain multi-event segments with separated signals available. We pay attention to re-mixing only segments recorded in the same room and keep the channel order consistent. One problem of this method is that the groundtruth separated events created this way still have some additive noise background. This noise background is distinct for both events and may be a source of overfitting. Nevertheless, we use this method to train a separation network

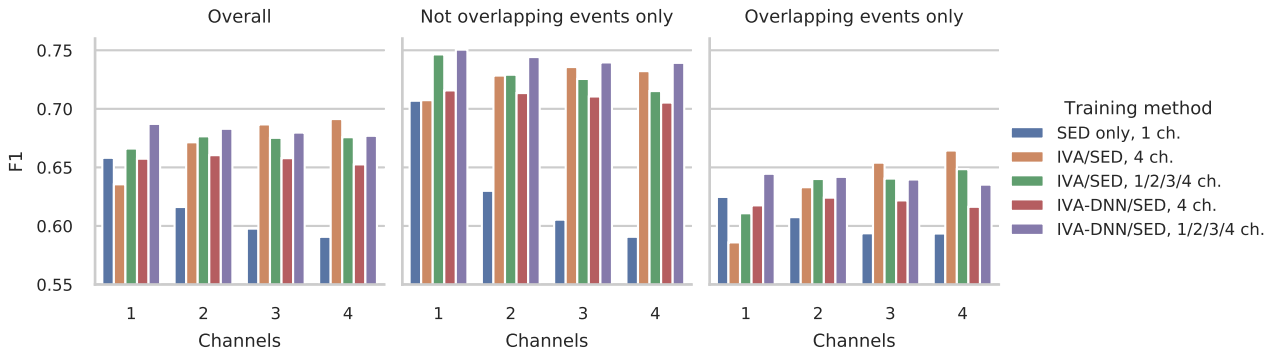

Fig. 2. F1 score on the test split of the dataset. From left to right, overall performance, on non-overlapping events only, and on overlapping events only.

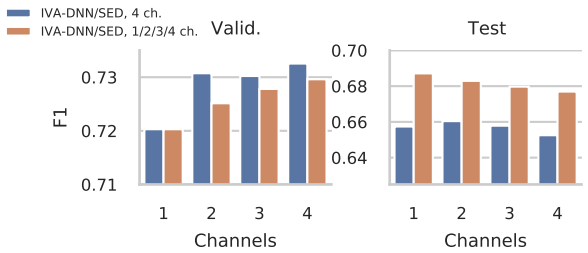

Fig. 3. Comparison of F1 score on the validation and test split for the DNN based separation.

with the DNN described in Fig. 1. The input spectrogram is transformed to 32 channels mel-spectrogram, which is then fed through two layers of BLSTM. The output of the BLSTM is brought back to the original spectrogram size through a single layer of 1×1 transposed convolution. As in [31], we train on two channel data only.

IV. Experiments

In our experiments, we compare the classification accuracy, via the F1 score, after training with the different strategies described in the previous section. We use the development dataset of the DCASE2020 SELD challenge [21]. It consists of 600 samples, each $1\,\mathrm{min}$ long, sampled at $24\,\mathrm{kHz}$, and with 4 channels. They are divided in 6 equal-sized folds: 4 for training, 1 for validation, and 1 for test. We use the training splits to adjust the model weights, the validation for the hyperparameters, and the test fold only a single time to obtain the results reported in this section. All samples are divided into $7.5\,\mathrm{s}$ long segments prior to processing. We further create single, two, and, three channel segments by taking all possible distinct subsets of channels as described in Section III-B. This multiplies the number of samples by 4, 6, and 4, respectively, compared to their original number. We use an STFT with a 1024-point Hamming window and half overlap.

We compare five different training regimes. **SED only, 1ch.**: We train the SED model on the single-channel dataset. IVA uses the time-varying Gauss source model [33]. **IVA/SED, 4ch.**: We train the SED model on the 4-channel dataset. IVA uses the time-varying Gauss source model [33]. **IVA/SED, 1/2/3/4ch.**: Same as the previous one, but we mix all the training data available. **IVA-DNN/SED, 4ch.**: We train the SED model on the 4-channel dataset. IVA uses the source model trained as described in Section III-C. **IVA-DNN/SED, 1/2/3/4ch.**: Same as the previous one, but we mix all the training data available. We use the F1 score, i.e., the harmonic mean of precision and recall [34], as a measure of the classification performance. The classification threshold applied to the event probability vector to obtain the final decision is chosen as the one that yields the best performance on the validation data. All experiments are performed on a Linux workstation with an Nvidia V100 GPU.

A. Results

Fig. 2 shows the F1-score of all the trained models on the test data as a function of the number of channels. We further break down the results between segments that do not contain overlapping events and those that do, (at most two in the dataset).

First, we can verify that *SED only, 1 ch.* performs best on single channel data, and that performance is actually hurt by the application of IVA with more channels. Second, we see that we can recover from this by training on the output of separation. *IVA/SED, 4 ch.*, although trained only on 4 channel data yields progressively better performance going from 2 to 4 channels at test time. For 4 channels, this is the best method overall. Without surprise, we can see that most of the gain comes from much improved performance on overlapping events. By training on a combination of number of channels in *IVA/SED, 1/2/3/4 ch.*, we actually surpass performance of the single channel training for single channel test data. While, this model still gives slightly better performance with more channels, the performance for 3 and 4 channels is worse than the previous model.

Surprisingly, using IVA with the pre-trained DNN model, and training with all available channel numbers yields the best single channel performance overall. However, the performance does not increase with the number of channels as expected. We conjecture that this is due to an overfit of the separation model on the training data. As supporting evidence, we compare the validation and test results in Fig. 3. There, we observe

that using more channels indeed helps on validation, but not test data. This may be due to the confounding background noise present in the available reference signals for training the separation model. However, it seems that using this method yields an effective data augmentation strategy that exploits the output of the (failed) separation step as new training data.

V. Conclusion

The preliminary results in this paper indicate that multichannel recordings are beneficial in several ways for SED. First, they can be used to improve the detection accuracy at test time, i.e., using more channels provides better detection performance. We explored different ways of combining the number of channels at training time to balance the performance at test time. This shows that it is possible to train a single model to be deployed on devices with varying number of channels. Second, through a failed attempt at training a tailored separation model for SED, we discovered that multichannel recordings may be used for effective data augmentation. In this case, we found large improvement at test time on single channel recordings when including multichannel recordings during training. However, there is room to improve the training scheme when using separation in order to obtain the best performance with any number of channels. In addition, we believe the potential for data augmentation may lead to large performance improvements in the future. We hope to fill out these gaps in future work.

References

[1] D. Bragg, N. Huynh, and R. E. Ladner, "A personalizable mobile sound detector app design for deaf and hard-of-hearing users," in *Proc. ACM ASSETS*, New York, NY, USA, Oct. 2016, pp. 3–13.

[2] R. I. Damper and M. D. Evans, "A multifunction domestic alert system for the deaf-blind," *IEEE Trans. Rehabilitation Eng.*, vol. 3, no. 4, pp. 354–359, Dec. 1995.

[3] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *2005 IEEE Int. Conf. Multimedia Expo*, Amsterdam, NL, Jul. 2005, pp. 4–pp.

[4] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, Oct. 2005, pp. 158–161.

[5] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–46, May 2016.

[6] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 6, pp. 992–1006, Apr. 2019.

[7] A. Gorin, N. Makhazhanov, and N. Shmyrev, "DCASE 2016 sound event detection system based on convolutional neural network," in *IEEE AASP Challenge: DCASE2016*, 2016.

[8] N. Takahashi, M. Gygli, B. Pfister, and L. V. Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Proc. ISCA INTERSPEECH*, San Francisco, CA, USA, Sep. 2016, pp. 2982–2986.

[9] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE ICASSP*, Shanghai, CN, Mar. 2016, pp. 6440–6444.

[10] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 11, pp. 2059–2070, Nov. 2017.

[11] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[12] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE ICASSP*, New Orleans, LA, USA, May 2017, pp. 771–775.

[13] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly-supervised sound event detection with self-attention," in *Proc. IEEE ICASSP*, Barcelona, ES, May 2020, pp. 66–70.

[14] ——, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proc. DCASE2020 Workshop*, Tokyo, JP, Nov. 2020.

[15] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[16] T. Komatsu, Y. Senda, and R. Kondo, "Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation," in *Proc. IEEE ICASSP*, Shanghai, CN, Mar. 2016, pp. 2259–2263.

[17] M. Brandstein and D. Ward, Eds., *Microphone Arrays*, ser. Signal Processing Techniques and Applications. Springer, Dec. 2010.

[18] H. L. Van Trees, *Optimum Array Processing*. New York, USA: John Wiley & Sons, Inc., Mar. 2002.

[19] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[20] S. Makino, Ed., *Audio Source Separation*, ser. Signals and Communication Technology. Cham: Springer International Publishing, 2018.

[21] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. DCASE2020 Workshop*, Tokyo, JP, Nov. 2020, pp. 165–169.

[22] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Apr. 2019.

[23] Q. Wang, H. Wu, Z. Jing, F. Ma, Y. Fang, Y. Wang, and T. Chen, "The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge," Proc. DCASE2020 Workshop, Tokyo, JP, Tech. Rep., Nov. 2020.

[24] T. N. Tho Nguyen, D. L. Jones, and W. S. Gan, "Ensemble of sequence matching networks for dynamic sound event localization, detection, and tracking," in *Proc. DCASE2020 Workshop*, Tokyo, JP, Nov. 2020, pp. 120–124.

[25] T. Komatsu, M. Togami, and T. Takahashi, "Sound event localization and detection using convolutional recurrent neural networks and gated linear units," in *Proc. IEEE EUSIPCO*, Amsterdam, NL, Jan. 2021.

[26] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Oxford, UK: Academic Press/Elsevier, 2010.

[27] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *ASIACRYPT 2016*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 601–608.

[28] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 70–79, Dec. 2006.

[29] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, Oct. 2011, pp. 189–192.

[30] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, Barcelona, ES, May 2020, pp. 236–240.

[31] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. IEEE ICASSP*, Toronto, CA, Jun. 2021, accepted.

[32] K. Lange, *MM optimization algorithms*. SIAM, 2016.

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

[34] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, Jun. 2016.