INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-channel separation of dynamic speech and sound events

Takuya Fujimura[1,2], *Robin Scheibler*[2]

[1]Nagoya University, Japan
[2]LINE Coporation, Japan

fujimura.takuya@g.sp.m.is.nagoya-u.ac.jp

Abstract

We propose a multi-channel separation method for moving sound sources. We build upon a recent beamformer for a moving speaker using attention-based tracking. This method uses an attention mechanism to compute the time-varying spatial statistics which enables tracking the moving source. While this prior work aimed to extract a single target source, we simultaneously estimate multiple sources. Our main technical contribution is to introduce attention-based tracking into the iterative source steering algorithm for independent vector analysis (IVA), enabling joint estimation of multiple sources. We experimentally show that the proposed method greatly improves the separation performance for moving speakers, including an absolute reduction of 27.2% in word error rate compared to time-invariant IVA. In addition, we demonstrate that the proposed method is effective as a pre-processing for sound event detection, showing an improvement in F1 scores of up to 4.7% in real recordings.
Index Terms: speech separation, moving source, independent vector analysis, self-attention network

1. Introduction

Sound source separation is a technique to estimate individual sources from their mixture [1]. Its variants can be broadly classified into single- and multi-channel methods. Single-channel methods achieve very high performance through the supervised learning of deep neural networks (DNNs) [2]. However, they suffer from distortion, and the domain shift between training and evaluation is often problematic [3]. On the other hand, multi-channel methods add an extra spatial dimension which enables low distortion and robust performance. But they require multiple microphones. Beamforming and independent vector analysis (IVA) are two examples of multi-channel methods. Beamforming targets a single source [4], while IVA simultaneously estimates beamformers for multiple sources assumed statistically independent [5, 6]. While traditionally limited by the accuracy of the estimated source statistics, using DNNs for the task has recently led to a breakthrough in performance [7, 8].

These sound source separation techniques are used as pre-processing for various applications, e.g., automatic speech recognition (ASR). For example, beamformer has been shown to reliably improve the performance of ASR due to its distortionless processing [9]. Beyond speech, separation has been recently used as pre-processing for sound event detection (SED) [10, 11, 12, 13, 14]. For single-channel SED, separation has been shown to be useful [11] but the performance is likely adversely impacted by domain shift. The unsupervised learning of separation was shown to be helpful for bird classification [10]. Similarly, with multi-channel recordings, initial work has shown the effectiveness of IVA as a pre-processing method [12, 13]. Nevertheless, the effectiveness of multi-channel methods based on beamforming and IVA for speech and SED has been so far mostly demonstrated for static sources. This is a serious challenge to their practical applicability.

To alleviate this limitation, Ochiai *et al*. have proposed to combine beamforming with self-attention-based tracking [15]. Generally, the spatial covariance matrices (SCMs) needed for the computation of the beamforming weights are obtained by time-averaging the instantaneous SCMs (ISCMs) over multiple frames, assuming that the sources are not moving. This results in a time-invariant demixing matrix. Instead, they propose to use *attention weights* that indicate which ISCM to focus on for estimating the demixing matrix at each frame. Attention weights are estimated from ISCMs through the self-attention network. ISCMs capture the local spatial information of the source and the attention weights combine it into robust instantaneous spatial statistics. The effectiveness of this approach has been shown for speech enhancement, but its extensions to the separation of multiple sources have yet to be investigated.

Our contributions in this paper are as follows. 1) Based on the success of DNN-supported IVA and self-attention-based tracking, we propose multi-channel separation methods for moving sound sources. We propose both a straightforward extension of [15] to multiple targets, and a time-varying IVA that introduces the attention mechanism into the low-complexity iterative source steering (ISS) algorithm [16]. 2) We propose an efficient architecture for estimating the attention weights from simple spatial features with lower dimensionality than the ISCMs used in [15]. 3) We demonstrate the effectiveness of the approach both for speech separation and sound event detection. Experimental results showed that the proposed methods achieved separation even for moving sources, unlike the conventional time-invariant approach, and the proposed separation improved SED performance by properly extracting moving events such as footsteps.

2. Backgrounds

We model the observed mixture in the short-time Fourier transform (STFT) domain represented as,

$$\boldsymbol{x}_{ft} = \boldsymbol{A}_{ft}\boldsymbol{s}_{ft} + \boldsymbol{b}_{ft}, \tag{1}$$

where $\boldsymbol{A}_{ft} \in \mathbb{C}^{M\times K}$ is the mixing matrix, $\boldsymbol{s}_{ft}$ is the clean sources, $\boldsymbol{b}_{ft}$ is the background noise, and M and K are the number of microphones and sources, respectively. Notations $f = 1, ..., F$ and $t = 1, ..., T$ denote the frequency bin and the time frame index, respectively. Assuming that the source is not moving, the mixing matrix becomes time-invariant, i.e., $\boldsymbol{A}_{ft} = \boldsymbol{A}_f$. Hereafter, $\boldsymbol{A}^\top$ and $\boldsymbol{A}^\mathsf{H}$ denote the transpose and conjugate transpose of $\boldsymbol{A}$, respectively.

10.21437/Interspeech.2023-164

2.1. Independent vector analysis

IVA is a blind source separation method. Assuming static sources and the determined case, where $M = K$, we obtain the separated sources $\boldsymbol{y}_{ft}$ using the time-invariant demixing matrix $\boldsymbol{W}_f^{\mathrm{IVA}} \in \mathbb{C}^{M\times M}$ as,

$$\boldsymbol{y}_{ft} = \boldsymbol{W}_f^{\mathrm{IVA}} \boldsymbol{x}_{ft}. \tag{2}$$

The demixing matrix is estimated by maximizing the log-likelihood of the observed signals, assuming independence of the sources and a statistical distribution for individual sources. For the optimization, we use ISS, an efficient iterative method based on majorization-minimization [16]. ISS updates the demixing matrix sequentially for $m = 1$ to M [16]:

$$\boldsymbol{W}_f^{\mathrm{IVA}} \leftarrow \boldsymbol{W}_f^{\mathrm{IVA}} - \boldsymbol{v}_{kf}(\boldsymbol{w}_{kf}^{\mathrm{IVA}})^{\mathsf{H}}, \tag{3}$$

$$u_{mkft} = r_{mft} y_{mft}(y_{kft})^*, \quad d_{mkft} = r_{mft}|y_{kft}|^2, \tag{4}$$

$$v_{mkf} = \begin{cases} \frac{\sum_t u_{mkft}}{\sum_t d_{mkft}} & \text{if } m \neq k, \\ 1 - \left(\sum_t d_{mkft}\right)^{-1/2} & \text{if } m = k, \end{cases} \tag{5}$$

where $\boldsymbol{v}_{kf} = [v_{1kf}, ..., v_{Mkf}]^\top$, $(\boldsymbol{w}_{kf}^{\mathrm{IVA}})^{\mathsf{H}}$ is the kth row of $\boldsymbol{W}_f^{\mathrm{IVA}}$, and $r_{mft} = \varphi_{ft}(\boldsymbol{Y}_m)$ with $(\boldsymbol{Y}_m)_{ft} = y_{mft}$. Albeit derived from the source model, the function $\varphi(\cdot)_{ft}$ can be interpreted as masking the target source to estimate statistics of noise and interference [16]. This makes it a good target to be replaced by a DNN, which has been shown to significantly improve the performance [8].

We note that a method derived from IVA for moving sound sources with the time-invariant filter that is robust to the movement has been proposed [17]. However, the moving area is restricted due to the time-invariant filter, and no solution has been presented yet for the general case.

2.2. Beamforming of one moving source

The beamformer aims to extract one source (i.e., $K = 1$) as,

$$y_{fn} = (\boldsymbol{w}_f)^{\mathsf{H}} \boldsymbol{x}_{fn}. \tag{6}$$

We can compute the beamformer weight vector $\boldsymbol{w}_f \in \mathbb{C}^M$ using the SCMs. For example, the minimum variance distortionless response (MVDR) beamformer weights are given by,

$$\boldsymbol{w}_{ft}^{\mathrm{MVDR}} = \frac{(\boldsymbol{\Phi}_{ft}^{\mathrm{N}})^{-1}\boldsymbol{\Phi}_{ft}^{\mathrm{S}}}{\mathrm{Tr}((\boldsymbol{\Phi}_{ft}^{\mathrm{N}})^{-1}\boldsymbol{\Phi}_{ft}^{\mathrm{S}})}\mathbf{u}, \tag{7}$$

where $\boldsymbol{\Phi}_{ft}^{\mathrm{S}} \in \mathbb{C}^{M\times M}$ and $\boldsymbol{\Phi}_{ft}^{\mathrm{N}} \in \mathbb{C}^{M\times M}$ are SCMs of target source and noise, respectively. $\mathbf{u} \in \mathbb{R}^M$ is a one-hot vector selecting the reference microphone and $\mathrm{Tr}(\cdot)$ denote the trace of the matrix. These SCMs are estimated using the time-frequency (T-F) masks as follows [18, 19, 15]:

$$\boldsymbol{\Phi}_f^\nu = \sum\nolimits_t \boldsymbol{\Psi}_{ft}^\nu, \quad \text{with} \quad \boldsymbol{\Psi}_{ft}^\nu = \gamma_{ft}^\nu \boldsymbol{x}_{ft}\boldsymbol{x}_{ft}^{\mathsf{H}}, \tag{8}$$

where $\boldsymbol{\Psi}_{ft}^\nu$ is the estimated ISCM, $\gamma_{ft}^\nu \in [0, 1]$ is a T-F mask normalized so that $\sum_t \gamma_{ft}^\nu = 1$, and $\nu \in \{\mathrm{S}, \mathrm{N}\}$ are the indexes for the target source and noise, respectively. Although Eq. (7) indicates the time-varying beamformer coefficients, it actually becomes time-invariant due to the time averaging of the ISCMs. Therefore, the beamformer cannot handle the moving sources.

To relax this limitation, Ochiai *et al.* have proposed the time-varying beamformer with self-attention-based weighting [15]. They propose to use an attention-mechanism to produce a time-varying weights matrix $\boldsymbol{c}^\nu \in \mathbb{R}^{T\times T}$, and replace the time-invariant SCM of (8) by a time-varying SCM,

$$\boldsymbol{\Phi}_{ft}^\nu = \sum\nolimits_\tau c_{t\tau}^\nu \boldsymbol{\Psi}_{ft}^\nu, \tag{9}$$

where $c_{t\tau} = (\boldsymbol{c}^\nu)_{t,\tau}$. In [15], the attention weights are estimated from the ISCMs by a multi-layer self-attention network. Since the ISCMs capture the spatial information, high-quality estimates of the time-varying SCMs are obtained. The method not only dramatically improved the separation performance for moving sources, but also static ones.

3. Proposed method

We propose two multi-channel separation methods for moving sources. The first is the straightforward extension of [15] to multiple targets where masks and attention matrices are estimated for all M sources by the supporting DNN. The second is a novel extension of IVA that incorporates the self-attention-based weighting into ISS. We modify the ISS updates as follows to estimate a time-varying demixing matrix $\boldsymbol{W}_{fn}$:

$$\boldsymbol{W}_{ft}^{\mathrm{IVA}} \leftarrow \boldsymbol{W}_{ft}^{\mathrm{IVA}} - \boldsymbol{v}_{kft}(\boldsymbol{w}_{kft}^{\mathrm{IVA}})^{\mathsf{H}}, \tag{10}$$

$$v_{mkft} = \begin{cases} \frac{\sum_\tau c_{mt\tau} u_{mkf\tau}}{\sum_\tau c_{mt\tau} d_{mkf\tau}} & \text{if } m \neq k, \\ 1 - \left(\sum_\tau c_{mt\tau} d_{mkf\tau}\right)^{-1/2} & \text{if } m = k. \end{cases} \tag{11}$$

Conventional IVA has used all frames for the estimation of v_{mkf}, but it is not appropriate for moving sources. In the proposed IVA, we expect that self-attention-based weighting selects the useful frame for updating the demixing matrix.

In addition, we propose in the following sub-sections alternative architectures for the attention and mask modules.

3.1. Attention module

Using the ISCMs directly as input as in [15] scales quadratically with the number of microphones. Instead, we modify a tried-and-tested architecture proposed for SELD [13]. Specifically, we obtain the attention weights as $\boldsymbol{c}_m = \mathrm{AttModule}(\boldsymbol{r}_m \odot \boldsymbol{x})$ where $\boldsymbol{r}_m$ is the T-F mask of the IVA or MVDR. $\mathrm{AttModule}(\cdot)$ receives the input $\mathrm{x} \in \mathbb{C}^{M\times F\times T}$ and processes it as follows:

$$\mathbf{z}_0 = \mathrm{Mel}(\mathrm{Concat}(|\mathrm{x}|^2, \mathrm{SpaFeat}(\mathrm{x}))) \tag{12}$$

$$\mathbf{z}_1 = \mathrm{Conv}(10\log_{10}(\mathbf{z}_0)), \tag{13}$$

$$\boldsymbol{c}_m = \mathrm{SelfAtt}(\mathbf{z}_1), \tag{14}$$

where the $\mathrm{Mel}(\cdot)$ is mel-scale transform with 128 filterbanks and $\mathrm{SpaFeat}(\cdot)$ extracts the spatial features with D dimensions such as the intensity vector [20, 21, 22]. $\mathrm{Conv}(\cdot)$ transforms $\mathbf{z}_0 \in \mathbb{R}^{M+D\times 128\times T}$ to $\mathbf{z}_1 \in \mathbb{R}^{T\times 128}$ through convolution layers. $\mathrm{SelfAtt}(\cdot)$ calculates the self-attention using linear and softmax layers through a Transformer encoder where the number of heads is four and the dimension of a feed-forward network is 1000. Preliminary experiments showed that SELD features work as well as ISCMs, despite the smaller dimension.

3.2. Mask module

In prior work for DNN-supported IVA, the mask is estimated independently for all source estimates [8]. In the early iterations of ISS, such an approach cannot effectively estimate the masks.

Table 1: *Average SDR (*dB*), PESQ, STOI, and WER (%) of the separated signals from the test set. The average WER of the reverberant target sources on moving-0, moving-1, and moving-2 were 10.7, 10.7, and 11.0, respectively.* * *indicates that it is the reference score because the oracle masks were also used during the evaluation.*

Method	moving-0				moving-1				moving-2			
	SDR↑	PESQ↑	STOI↑	WER↓	SDR↑	PESQ↑	STOI↑	WER↓	SDR↑	PESQ↑	STOI↑	WER↓
Mixture	-0.15	1.28	0.70	72.2	-0.15	1.28	0.71	72.3	-0.14	1.28	0.71	74.0
ATT-MVDR	9.61	1.95	0.87	13.9	6.65	1.67	0.82	20.8	4.83	1.50	0.77	34.3
oracle mask*	11.56	2.12	0.90	11.6	8.80	1.87	0.87	13.9	7.18	**1.71**	**0.84**	**19.5**
TIV-IVA	10.65	1.95	0.89	11.6	4.06	1.55	0.79	20.8	-0.02	1.27	0.67	54.9
ONL-IVA	5.14	1.49	0.79	19.1	1.97	1.33	0.72	36.1	-0.20	1.24	0.66	60.1
BLK-IVA	8.84	1.79	0.87	14.2	4.49	1.49	0.79	20.5	1.86	1.32	0.71	44.0
ATT-IVA	**13.54**	**2.36**	**0.93**	**11.3**	**10.78**	**2.12**	**0.90**	**13.1**	**7.65**	**1.70**	**0.83**	**27.7**

To improve the convergence speed of ISS iterations, we use instead a network that jointly estimates the mask for all source estimates. Nevertheless, our mask module is almost the same as the conventional one using the GLU block layers [8]. For the joint estimation, the input log-magnitude spectrograms are stacked along the frequency axis and the dimension is reduced to F from MF by linear projection. Then, this is followed by the GLU Blocks and a transposed convolution layer as [8] and we split the output to obtain M T-F masks.

For MVDR, we adopt a mask module of bidirectional long short-term memory layers with 400 hidden units [23].

4. Experiments

We conducted two separate experiments for speech and sound events. Beyond separation, we assess the improvement of the downstream task metrics for ASR and SED, respectively.

4.1. Speech separation

4.1.1. Experimental setup

Dataset: We generate a dataset of noisy and reverberant mixtures of two speech sources. The dataset is replicated three times: with two static sources (`moving-0`), one static and one dynamic sources (`moving-1`), and two dynamic sources (`moving-2`). The speech is from WSJ0 [24] and WSJ1 [25] datasets with pairs and relative mixing ratios ($-5\,\mathrm{dB}$ to $5\,\mathrm{dB}$) identical to the WSJ1_2mix dataset [26]. The noise is from the CHiME3 [27] dataset, mixed with SNR from $10\,\mathrm{dB}$ to $30\,\mathrm{dB}$ relative to the reverberant speech mixtures. The sampling rate is $16\,\mathrm{kHz}$. Reverberation is simulated using the randomized image source method [28] implemented in pyroomacoustics [29]. The scene geometry loosely follows [15]. The rectangular rooms dimensions are random between $3\,\mathrm{m}$ to $8\,\mathrm{m}$. The two sources and the center of the microphone array are at least $50\,\mathrm{cm}$ from walls. The two microphones correspond to channels 3 and 6 of the array of CHiME3. The x coordinate of the array center is within $0.5\,\mathrm{m}$ to $1.5\,\mathrm{m}$. Accordingly, sources have x coordinate larger than $1.5\,\mathrm{m}$. For moving sources, start and end locations are sampled uniformly at random from the allowed volume. The trajectory is discretized on 20 points uniformly sampled from the line segment. The source speed is determined by the length of the segment and of the speech.

Comparison methods: We compared the proposed attention-based IVA (ATT-IVA) and MVDR (ATT-MVDR) with time-invariant (TIV-IVA), online (ONL-IVA), and blockwise IVA (BLK-IVA). TIV-IVA used $\boldsymbol{c}_{mt} = 1/T$ and, ONL-IVA and BLK-IVA obtained it as in [15] with the forgetting factor of 0.999 and the block size of 50, respectively, where all IVA methods use the mask module described in Sec. 3.2. In addition to joint optimization of mask and attention of ATT-MVDR, we also train the attention only with an oracle mask model [30] to understand the performance limits. For TIV-IVA, only the static dataset `moving-0` was used for the training and validation.

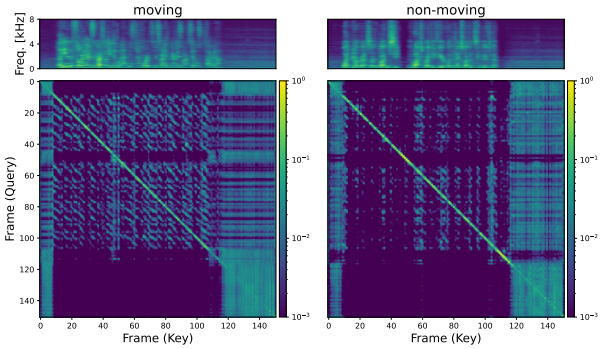

Figure 1: *Visualization of the attention weight. The colormap is in logarithmic scale to visualize small values.*

Training details: Unlike previous work [15], we simultaneously optimized the attention and mask modules with the source-aggregated signal-to-distortion ratio (SA-SDR) loss [31]. Since we are in the determined case, we can use $\boldsymbol{A} = \boldsymbol{W}^{-1}$ to obtain the source steering vectors and reconstruct M channels separated signals. The optimizer was Adam with a learning rate 0.0001 and linear warm-up over the first 10,000 steps. The minibatch size was eight and the training samples were trimmed or zero-padded to seven seconds. We trained for at least 150 epochs and up to 420 epochs to select the best checkpoints. The STFT used window and shift sizes of 4096 and 1024, respectively, and a Hann window. The number of ISS iterations was ten for TIV-IVA and five for the others. For $\mathrm{SpaFeat}(\cdot)$, we used the frequency-normalized interchannel phase differences [32, 33].

4.1.2. Experimental Results

Table 1 summarizes the evaluation results where the metrics are SDR, perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). In addition, we evaluate the word error rate (WER) with the Whisper `large-v2` model [34]. First, we can confirm that we succeeded in jointly optimizing the attention and mask modules of ATT-MVDR and ATT-IVA, unlike the previous work. ATT-IVA achieved the best performance under most conditions, even when compared to ATT-MVDR with oracle masks. As in previous reports [15], introducing attention weight improved the performance not only for moving sources but also for non-moving sources. Certainly, using attention allows to dynamically adapt to varying noise.

Table 2: *Average SDR (↑) for sound event separation on the synthetic validation set.*

Method	Mixture	TIV-IVA	**ATT-IVA**
SDR ↑	-6.04	-4.13	**0.71**

Table 3: *Average macro-F1 score on the validation split of the STARSS22 dataset [35] of real scenes. The ⋆ indicates that Classwise uses some oracle data and is provided for reference.*

MIX	TIV	ATT	TIV+ATT	Classwise⋆
0.5559	0.5681	0.5706	**0.5793**	**0.6026**

Figure 1 shows an example of spectrograms and attention weights from `moving-1`. We can see that the weights are concentrated on the diagonal, i.e., local information, especially when the interference source is moving. However, a large number of small magnitude off-diagonal components are also used, which likely helps stabilizing the SCM computation. It is reasonable that the weights are more uniformly distributed when the interference source is not moving since multiple frames have similar spatial characteristics and can be used for averaging.

4.2. SED with separation

4.2.1. Experimental setup

Dataset: The sound event separator is trained on a dataset of simulated sound event mixtures created following the methodology of the DCASE 2022 SELD baseline [35]. The original dataset did not contain sufficiently many moving sources and was modified as follows. The maximum overlap count is increased to four. Events from all classes move with probability 0.75. We added sounds from non-target classes to allow separation of interfering sounds. We generated 36.7 h and 3.3 h of training and validation data, respectively. For the training of the SED models, we used the official DCASE 2022 Task 3 baseline training and validation sets [35]. Testing was conducted on the validation set since the official test set has not yet been released. The DOA annotations are discarded since we focus on the SED task only. We use the first-order ambisonics format with four channels sampled at 24 kHz. The number of classes is thirteen.
Model structure and comparison methods: We built the SED models with and without separated signals. The models receive log-mel spectrograms of the single-channel signals. When we use only the mixture, we obtain an event probability vector through $\mathrm{Conv}(\cdot)$, eight conformer layers with a kernel size of 7, a linear layer, and a sigmoid function. When we use separation, two sets of $\mathrm{Conv}(\cdot)$ and conformer layers are prepared to extract the feature of the mixture and separated signals, respectively. Then, a feature of the mixture is added to each of the features obtained from M separated signals. From the features, we obtain M event probability vectors and they are combined by the max operation. We used the max operation because only one of the M predictions may detect the event if each event was perfectly separated [12]. We built several SED models using separation, but this late fusion approach was the best as in [11].
Training and evaluation details: We trained the TIV- and ATT-IVA for 100 epochs with five ISS iterations. The window and shift size of the STFT were 2048 and 512 samples. For ATT-IVA, we used the intensity vector as $\mathrm{SpaFeat}(\cdot)$ and set the weight decay to 0.0001 to prevent overfitting. We trained the SED models for 1,000 epochs using binary cross-entropy loss, Adam optimizer with a learning rate of 0.001, and SpecAugment [36]. The minibatch size was 128. We do a learning rate warm-up over the first 10,000 steps. For the STFT, the frame, window, and shift sizes were set to 1024, 600, and 240 samples, respectively. For the evaluation, we averaged the predictions of the best five models.

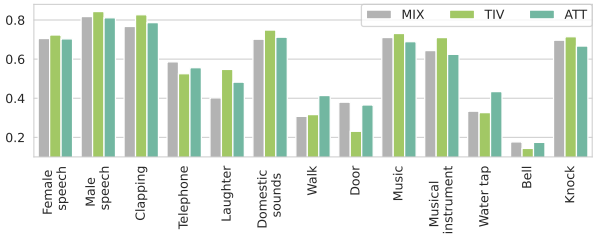

Figure 2: *Classwise analysis of the F1 scores*

4.2.2. Experimental Results

Table 2 shows the SDR scores of the signals separated by TIV-IVA and ATT-IVA on the dataset of synthetic event mixtures. The absolute SDR is low, but TIV- and ATT-IVA provide 2 and 6 dB improvement, respectively, over the mixture SDR.

Next, we evaluated the SED performance. Table 3 shows the macro-F1 score. MIX used only the mixture, while TIV and ATT used the mixture and the separated signals by TIV-IVA and ATT-IVA, respectively. TIV+ATT used the average of the output of ten models, five TIV and five ATT. From the Table 3, we can see that the separation is effective since TIV and ATT outperform MIX. When breaking down the performance by class, shown in Figure 2, we observed that TIV and ATT have very complementary strengths and weaknesses. As one would expect, ATT performed very well for *Walk*, which clearly involves movement. Surprisingly, ATT also achieved high performance for *Water tap*. Informal listening tests revealed that several such samples also feature loud music that only ATT was able to isolate from the water sounds (see *Example1* of the SED experiments at our demo site[1]). On the other hand, TIV performed best for fairly static sources such as speech. This motivated the combination of both classifiers (ATT+TIV) by averaging, which lead to a boost of close to 1 %. Going yet further, one could choose to use either or both classifiers on a per class basis. We test this hypothesis by only using the best of MIX/TIV/ATT as shown in Figure 2 and achieve F1 score of 0.60 , an absolute improvement of 4.7 % over using only the mixture. However, this result should be treated as indicative only since we used the validation data to choose the classifiers to use. Regardless, combining different separation approaches is a promising direction.

5. Conclusion

We proposed two multi-channel separation methods for moving sources. ATT-MVDR is a straightforward extension of the conventional MVDR for moving speakers to multiple targets. ATT-IVA introduced attention weights to the ISS algorithm. In the evaluation of the speech separation, ATT-IVA achieved high performance including situations where both two speakers were moving. For SED, the performance was strongly dependent on the class, and combining time invariant and varying methods brought large improvements. Overall, we found IVA easier to train and performing better than MVDR-based separation, which we attribute to the joint formulation of IVA and the good numerical properties of ISS. The combination of ATT and TIV is a promising research line. Furthermore, the rapid progress of all neural approaches should be taken into account [37].

[1] `http://www.robinscheibler.org/interspeech2023-moving-iva-samples`

6. References

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

[2] Z.-Q. Wang *et al.*, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," Sep. 2022, arXiv:2209.03952.

[3] T. Cord-Landwehr *et al.*, "Monaural source separation: From anechoic to reverberant environments," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[4] H. L. Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2002.

[5] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *ICA 2006*, ser. Lecture Notes on Computer Science. Springer Berlin Heidelberg, Jan. 2006, vol. 3889, pp. 165–172.

[6] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *ICA 2006*, ser. Lecture Notes on Computer Science. Springer Berlin Heidelberg, Jan. 2006, vol. 3889, pp. 601–608.

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

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

[9] R. Haeb-Umbach *et al.*, "Far-Field Automatic Speech Recognition," *Proc. IEEE*, vol. 109, no. 2, pp. 124–148, Feb. 2021.

[10] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *Proc. IEEE ICASSP*, Singapore, SG, May 2022, pp. 636–640.

[11] N. Turpault *et al.*, "Improving sound event detection in domestic environments using sound separation," in *Proc. DCASE*, Tokyo, JP, Nov. 2022.

[12] R. Scheibler, T. Komatsu, and M. Togami, "Multichannel separation and classification of sound events," in *Proc. EUSIPCO*, Dublin, IRL, Aug. 2021, pp. 1035–1039.

[13] R. Scheibler, T. Komatsu, Y. Fujita, and M. Hentschel, "Sound event localization and detection with pre-trained audio spectrogram transformer and multichannel seperation network," in *Proc. DCASE*, Nancy, France, Nov. 2022.

[14] I. Kavalerov *et al.*, "Universal sound separation," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 175–179.

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

[16] 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.

[17] Z. Koldovskỳ, V. Kautskỳ, P. Tichavskỳ, J. Čmejla, and J. Málek, "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2158–2173, 2021.

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

[19] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, San Francisco, USA, Sep. 2016, pp. 1981–1985.

[20] D. P. Jarrett, E. A. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. EUSIPCO*, Aalborg, DK, 2010, pp. 442–446.

[21] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D localization of multiple sound sources with intensity vector estimates in single source zones," in *Proc. EUSIPCO*, Nice, FR, 2015, pp. 1556–1560.

[22] M. Yasuda, Y. Koizumi, S. Saito, H. Uematsu, and K. Imoto, "Sound event localization based on sound intensity vector refined by DNN-based denoising and source separation," in *Proc. IEEE ICASSP*, Barcelona, ES, May 2020, pp. 651–655.

[23] C. Boeddeker *et al.*, "Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation," in *Proc. IEEE ICASSP*, Toronto, CA, Jun. 2021, pp. 8428–8432.

[24] J. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete LDC93S6A*, Linguistic Data Consortium, Philadelphia, 1993, web Download.

[25] Linguistic Data Consortium, and NIST Multimodal Information Group, *CSR-II (WSJ1) Complete LDC94S13A*, Linguistic Data Consortium, Philadelphia, 1994, web Download.

[26] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. ACL*, Melbourne, Australia, Jul. 2018, pp. 2620–2630.

[27] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, Scottsdale, USA, Dec. 2015, pp. 504–511.

[28] E. De Sena, N. Antonello, M. Moonen, and T. van Waterschoot, "On the modeling of rectangular geometries in room acoustic simulations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 4, pp. 774–786, Apr. 2015.

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

[30] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE ICASSP*, Brisbane, AUS, Apr. 2015, pp. 708–712.

[31] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "SA-SDR: A novel loss function for separation of meeting style data," in *Proc. IEEE ICASSP*, Singapore, SG, May 2022, pp. 6022–6026.

[32] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *Proc. IEEE ICASSP*, Singapore, SG, May 2022, pp. 716–720.

[33] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[34] A. Radford *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, arXiv:2212.04356 [cs, eess].

[35] A. Politis *et al.*, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. DCASE*, Nancy, FR, Nov. 2022.

[36] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, Graz, AUT, 2019, pp. 2613–2617.

[37] D. Marković, A. Défossez, and A. Richard, "Implicit neural spatial filtering for multichannel source separation in the waveform domain," in *Proc. Interspeech*, Incheon, KR, Sep. 2022, pp. 1806–1810.