AN END-TO-END DEEP LEARNING FRAMEWORK FOR MULTIPLE AUDIO SOURCE SEPARATION AND LOCALIZATION

Yu Chen, Bowen Liu, Zijian Zhang, Hun-Seok Kim

University of Michigan, USA

ABSTRACT

Sound source separation and localization for situational awareness enables a wide range of applications such as hearing enhancement and audio beam-forming. We present an end-to-end deep learning framework to separate and localize multiple audio sources from the mixture of multi-channels. The proposed framework jointly estimates the separated sources and their time difference of arrival (TDOA) at different microphones, then it obtains the direction-of-arrival (DOA) for each source. A new structure to reconstruct the mixed signal is introduced for joint optimization of source separation and TDOA estimation. In addition, a discriminator network is added during the training phase to further improve the separation quality. Experiment results demonstrate that the proposed method achieves state-of-the-art accuracy on source separation as well as DOA estimation.

***Index Terms*—** Multiple audio source localization, audio source separation, deep learning, discriminator

1. INTRODUCTION

Sound source localization is a problem to localize an acoustic source(s) in space using the captured audio from a microphone or microphone array by typically estimating the direction of arrival (DOA) or angle of arrival (AOA) of the source. In the case of multi-sources as in the *cocktail party problem*, a good source separation strategy is the premise for precise DOA/AOA estimation. Methods that combine both accurate source separation and localization can enable a wide range of practical applications including autonomous robot navigation, virtual reality (VR) headsets, and enhanced audio surveillance [1] in industrial facilities and critical environments.

In the early stage, source localization problems were mainly solved by analytical approaches [2, 3, 4] and their robustness is rather limited for practical applications. Recently, deep learning based methods [5, 6, 7, 8, 9] have shown superior performance over the traditional analytical approaches. Existing DNN based source localization methods typically take mixed signals as the input and produce DOA as the output for end-to-end model training without any interpretable intermediate information. Meanwhile, it has been shown that the audio source separation problem can be efficiently solved through deep learning based approaches even with a single microphone. State-of-the-art performance has been realized in different separation tasks such as speech [10, 11, 12], universal sound [13], and music [14].

Our approach is motivated by the observation that, for the multi-source localization problem, some intermediate information such as the separated source signals and time difference of arrival (TDOA) between microphones can be explicitly obtained and utilized to improve the overall system performance. Hence, we adapt an existing state-of-the-art source separation model and integrate it into the proposed framework with a TDOA and DOA estimation networks. They are jointly trained with a novel training method to improve both source separation quality and localization accuracy.

In the proposed approach, separated source signals and their TDOA on microphones are jointly estimated, then the DOA of each source is estimated from TDOA. Although we adopt a prior network for source separation, we improve its performance by introducing a companion TDOA estimation network and jointly training them with a new framework where one network assists the other to optimize a similarity loss between the reconstructed and original mixed signals. This new framework ensures the separated sources are realistic as measured by a discriminator and also sufficient to reproduce the original mixture when combined with the estimated TDOA. Our framework introduces a new multi-network structure to perform DOA estimation with superior/similar performance compared to the state-of-the-art, while producing interpretable intermediate information such as separated sources and TDOA on microphones. Compared to baseline cases where individual networks are trained in isolation, our joint-training scheme achieves superior performance for source separation as well as TDOA estimation because one network assists the other to reproduce realistic reconstructed mixtures during the localization process.

2. METHOD

Fig. 1 shows the overall datapath of the proposed scheme for joint source separation, TDOA estimation, and DOA estimation. The proposed model consists of three parts; a source separation network, TDOA estimation network, and DOA es-

This work was supported by Meta Platforms, Inc.

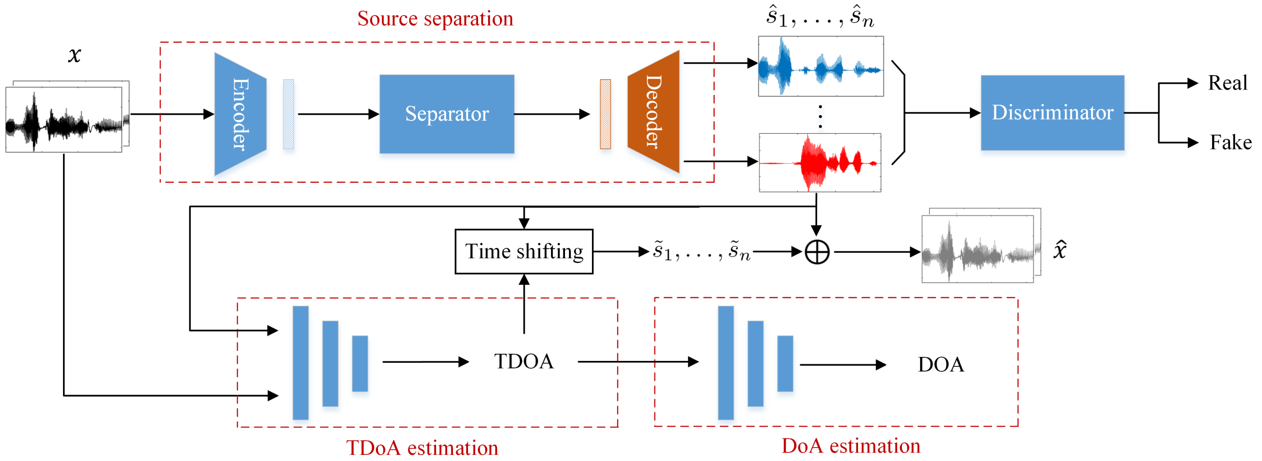

Fig. 1. Framework diagram. Source separation network estimates source signals $\hat{s}_1, \ldots, \hat{s}_n$ from multi-channel mixture x. TDOA information is estimated from x and $\hat{s}_1, \ldots, \hat{s}_n$ by TDOA estimation network. $\hat{x}$ is reconstructed by adding $\hat{s}_1, \ldots, \hat{s}_n$ and time shifted source signals $\tilde{s}_1, \ldots, \tilde{s}_n$. The reconstructed mixture $\hat{x}$ and discriminator are only used during the training.

timation network. TDOA information is estimated from the multi-channel mixtures, then sent to the DOA estimation network for localizing each source. The mixture is reconstructed using the separated sources and estimated TDOA, and we evaluate the similarity loss between the reconstructed $\hat{x}$ and original mixture x for joint optimization of source separation and TDOA estimation. In the meantime, a discriminator is added to improve the quality of separated signals.

2.1. Source separation

The source separation network extracts each audio source from multi-channel mixture. Since various audio source and speech separation models were previously investigated in the literature, we adapt some of the best models [10, 11] as the source separation network in our framework. As most of the separation models are designed for single channel (microphone), we choose models that perform separation in the latent domain, where encoder and decoder dimensions can be easily extended to multi-channel (microphone array) inputs and different number of sources. The separated audio quality is typically measured by scale-invariant signal to noise ratio (SI-SNR) [15], defined by

$$\text{SI-SNR}(s, \hat{s}) = 10 \log_{10} \frac{\|\alpha s\|^2}{\|\alpha s - \hat{s}\|^2}, \tag{1}$$

where $\alpha = \hat{s}^T s / \|s\|^2$, s is the target signal and $\hat{s}$ is the estimated signal. The separation network is trained to minimize the negative permutation-invariant SI-SNR [16], defined as

$$\mathcal{L}_{\text{sep}}(s^*, \hat{s}) = -\text{SI-SNR}(s^*, \hat{s}) = -10 \log_{10} \frac{\|\alpha s^*\|^2}{\|\alpha s^* - \hat{s}\|^2}, \tag{2}$$

where s^* denotes the permutation of the sources that maximizes SI-SNR.

2.2. TDOA estimation

We propose to simultaneously discriminate and localize N sound sources using the TDOA estimated with an array of K microphones. The TDOA ΔT_{ij} between a pair of microphones i and j regarding a certain source is defined as

$$\Delta T_{ij} = \frac{f_s}{c}(\|l_s - l_i\| - \|l_s - l_j\|), \tag{3}$$

where l_s is the location (coordinate) of the source, l_i and l_j are the locations of microphone i and j, f_s is the sampling frequency, c is the speed of sound, and $\|\cdot\|$ denotes Euclidean norm. There are total $K(K-1)/2$ different TDOAs but only $K-1$ are independent. So, we choose to use only ΔT_{1j} values for DOA estimation.

Instead of estimating TDOA by identifying the peak of the cross-correlation of two signals, we propose to use a TDOA estimation network, which can be easily integrated and jointly trained with the source separation network to build an end-to-end system. Each source signal received at microphone 1 is treated as the reference (non-shifted version) in the source separation stage. The TDOA estimation network uses the estimated reference signal together with the mixtures received at other microphones to estimate the TDOA between microphone pairs regarding the same source. Since the TDOA is discretized due to audio signal sampling and its maximum is limited by the configuration of the microphone array, we treat TDOA estimation as a classification problem where each class represents a possible TDOA in terms of sample index.

At this point, the source separation network and TDOA estimation network can be independently trained with their own loss functions. We use the original source signals for the initial training of the TDOA estimation network. After this preliminary independent training, we combine source separation and TDOA estimation networks for joint end-to-end training to improve both separation quality and TDOA estimation accuracy. For that, we define a similarity loss eq. (4) between the original mixture x and the reconstructed mixture $\hat{x}$ obtained by applying the estimated TDOA information to the separated source signals. In (4), K is the total number of channels (microphones) and $\langle \cdot, \cdot \rangle$ denotes inner product.

$$\mathcal{L}_{\text{sm}}(x, \hat{x}) = -\frac{1}{K} \sum_{i=1}^{K} \langle x_i^T, \hat{x}_i \rangle. \tag{4}$$

This loss is designed to make the reconstructed multi-channel mixture as close as possible to the original one. The main issue of applying this similarity loss to network training is the non-differentiable TDOA time-shift operation to reconstruct mixture signals. We mitigate this issue by treating softmax of the TDOA estimation output vector y_j as the channel impulse response and convolving it with the separated signal $\hat{s}_j$ to obtain a time shifted version $\tilde{s}_j$ of source j. Then the reconstructed mixture $\hat{x}_i$ of channel i can be obtained by

$$\hat{x}_i = \sum_{j=1}^{N} \tilde{s}_j = \sum_{j=1}^{N} \text{softmax}(y_j) * \hat{s}_j, \tag{5}$$

where N is the number of sources, and $*$ is convolution. This technique allows end-to-end training of source separation and TDOA estimation networks through back-propagation.

Inspired by the success of generative adversarial networks (GANs) [17], we adopt a discriminator network in our framework to distinguish estimated/separated source signals (fake samples) from original source signals (real samples). The discriminator is only used during training to improve the source separation quality. Our source separation network is treated as a ‘generator’ in GAN which is adversarially trained with the discriminator.

The total loss including the subjective discriminator loss, separation loss, TDOA estimation loss, and reconstruction loss is defined as

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{sep}}(s^*, \hat{s}) + \mathcal{L}_{\text{TDOA}} + \alpha \cdot \mathcal{L}_{\text{sm}}(x, \hat{x}) \\ + \beta \cdot \mathbb{E}_{\hat{s}}(\log(1 - D(\hat{s}))), \end{aligned} \tag{6}$$

where $\mathcal{L}_{\text{TDOA}}$ is the cross-entropy loss for TDOA classification, $D(\hat{s})$ is the discriminator output (estimated probability that $\hat{s}$ is real), and α and β are weights for loss terms.

2.3. DOA estimation

The DOA estimation network is connected to the TDOA estimation network, taking the estimated TDOA of each source as the input to obtain the the azimuth angle of sources regarding the microphone array. A simple multilayer perception model is sufficient for this regression problem. It is trained separately with the ground-truth TDOA and corresponding angles, then it is inserted to the system for the final DOA inference.

3. EXPERIMENT SETUP

3.1. System setup and dataset

In all our experiments, we use speech from different speakers as multiple sources sampled at 16 kHz. The microphone array contains $K = 4$ microphones placed in a square shape with 0.2-meter separation per dimension, which is reasonable for real-world applications such as mobile robots. The distance between sources and microphones are in the range of 1 to 3 meters, and all sources are placed with an azimuth from 0 to 180°. Since our goal is to estimate accurate azimuth angles, the microphone array and sources are restricted in a 2D plane.

Publicly available datasets for acoustic source localization and tracking, e.g. the LOCATA challenge [18], have limitations such as small amount of training data, preset number of sources, and restricted microphone array configurations. Hence, we generate our own datasets from LibriSpeech [19], a large-scale dataset with corpus of read English speech from hundreds of distinctive speakers. The training set is generated from `train-clean-100`, and speech mixtures are created by randomly mixing speech utterances from different speakers, similar to LibriMix [20]. To simulate microphone array outputs, we create multi-channel mixtures by mixing data using data augmentation process introduced in [21]. The distance and azimuth angle of each source are randomly selected. Then, individual speaker sources are TDOA-shifted and added to create the mixture received at each microphone. The testing set is generated from `test-clean` after removing idle periods greater than 0.5 seconds. We generate datasets with $N = 3$ and 4 sources for training and evaluation. The audio length is set to 2 seconds in all experiments.

3.2. Architecture and Training Details

Two state-of-the-art speech separation models, SuDoRM-RF [10] and DPTNet [11] are adapted to serve as our source separation network. These models perform separation in the latent domain, thus it is straightforward to adjust the encoder and decoder latent dimensions to accommodate $K = 4$ channel inputs and different numbers of sources ($N = 3$ or 4). For SuDoRM-RF, we use 16 U-ConvBlocks and ReLU as the mask activation function. For DPTNet, we shorten the model by reducing the number of Transformer blocks (IntraTransformer and InterTransformer) to 2. The optimizers and other hyperparameters are the same as in the original papers.

The TDOA estimation network is a 6-layer CNN with four 1D convolution layers followed by two fully connected layers. In our experiment, the maximum time-shift is less than 20 samples, thus our TDOA estimation network output has 41 (with positive and negative TDOA) classes in total. To speed up the joint training process, it is pre-trained with clean speech and time-shifted mixtures in our training set before the joint training with the separation network. The discriminator is a CNN with four 1D convolution layers, and it is trained

Table 1. Separation quality evaluation by SI-SNRi (dB) of the proposed framework. Sep., Recon., Disc. represent separator, reconstruction and discriminator respectively.

Separator	N	Sep. only	Sep. + Recon.	Sep + Recon. + Disc.
SuDoRM -RF	3	16.75	17.79	18.64
	4	13.06	14.25	14.57
DPTNet	3	14.69	16.92	17.37
	4	10.53	11.77	11.86

with the binary cross-entropy loss and Adam optimizer. We send noisy inputs to the discriminator to stabilize its training procedure, as introduced by Arjovsky and Bottou [22]. Source separation and TDOA estimation networks are jointly trained for 200 epochs before adding the discriminator for alternated adversarial training. We set the scalar weights $\alpha = 1$ and $\beta = 0.01$ for the loss function (6).

4. RESULTS

The separation quality is evaluated by SI-SNR improvement (SI-SNRi) in dB, which is the gain of SI-SNR on the separated signal over the mixture signal. We test SuDoRM-RF and DPTNet separator with $N = 3$ and 4 sources. SI-SNRi values are reported in three cases: separator only, separator with mixture reconstruction, and separator with mixture reconstruction and discriminator. The evaluation results on source separation quality are summarized in Table 1. The separation quality is improved by 1.3 – 2.7 dB with the proposed reconstruction structure and discriminator.

Source separation and TDOA estimation are jointly optimized to reduce distortion between the reconstructed and original mixture. The joint training with the proposed reconstruction loss improves the separation quality. SI-SNRi of separated sources further enhances as the discriminator loss is combined with the reconstruction loss to guide the separated signal to be more realistic speech from a single speaker. It is worth noting that the proposed framework is generalizable to other separation network structures.

Table 2 summarizes the evaluation results on TDOA and DOA estimation of the proposed framework. We report mean absolute error (MAE) on TDOA and DOA estimation. Besides, we also report the localization recall on DOA estimation, where the DOA output is considered true positive only if it is under a threshold of 5° absolute error. Compared to the baseline results without the reconstruction and discriminator, we achieve lower TDOA and DOA MAE as well as higher localization recall thanks to the improved separation quality and TDOA estimation accuracy.

Table 2. MAE of TDOA ($\mathrm{E}_{\mathrm{TDOA}}$) and DOA ($\mathrm{E}_{\mathrm{DOA}}$), and localization recall ($\mathrm{R}_{\mathrm{DOA}}$) of the proposed framework.

Separator	N	Without Recon. & Disc.			With Recon. & Disc.		
		$\mathrm{E}_{\mathrm{TDOA}}$	$\mathrm{E}_{\mathrm{DOA}}$	$\mathrm{R}_{\mathrm{DOA}}$	$\mathrm{E}_{\mathrm{TDOA}}$	$\mathrm{E}_{\mathrm{DOA}}$	$\mathrm{R}_{\mathrm{DOA}}$
SuDoRM -RF	3	22.4 ms	2.16°	93.9%	21.5 ms	2.10°	94.5%
	4	28.2 ms	3.25°	87.6%	24.4 ms	2.70°	91.4%
DPTNet	3	28.8 ms	3.00°	89.8%	24.1 ms	2.46°	92.7%
	4	44.4 ms	5.38°	81.9%	35.6 ms	3.96°	86.6%

Table 3. Percentage of non-anomalous frames and DOA error for proposed framework and SMESLP [9], I-IDIR-UCA [4], CHB [23] (reported in [9]).

Methods	Ours	SMESLP*	I-IDIR-UCA	CHB
Number of microphones	4	8	8	8
MAE	1.67°	2.05°	-	-
RMSE	3.01°	2.33°	4.1°	2.98°
Non-anomalous frames	98.1%	100%	60%	-

* SMESLP uses part of the testing sequence for fine-tuning and validation. Our method uses exclusive sequences for training, validation, and testing.

Finally, we compare our framework with state-of-the-art DOA estimation algorithms, SMESLP [9], I-IDIR-UCA [4] and CHB [23] in Table 3. The same recording sequence `seq37-3p-0001` from the AV16.3 corpus [24] of real indoor recording with 3 speakers is used for evaluation. We remove idle periods and remix the signal based on our microphone array setting since the original dataset uses a different microphone array. Unlike [9], our system contains only 4 microphones instead of 8. Results for SMESLP, I-IDIR-UCA and CHB are reported by [9]. Our method with 4 microphones outperform all three methods on mean absolute DOA error but get slightly worse root mean squared error (RMSE) compared to MSESLP and CHB with 8 microphones. The percentage of non-anomalous frames is comparable to SMESLP and better than the other two. Overall, the proposed framework achieves better or comparable performance with fewer number of microphones. In addition, our scheme produces interpretable intermediate outputs such as separated sources and their TDOAs, which are not available from other approaches.

5. CONCLUSION

We present an end-to-end deep learning framework for accurate source separation and localization in multi-source environments. By joint training of separation and TDOA estimation networks with a reconstruction structure and a discriminator network, the source separation quality as well as the TDOA estimation accuracy improves. Our experiment results confirm that the proposed framework achieves superior performance compared to the baseline. Our framework is generalizable to other source separation models and can be further improved with better separation models in the future.

6. REFERENCES

[1] M. Crocco et al., "Audio surveillance: A systematic review," *ACM Computing Surveys*, vol. 48, no. 4, pp. 1–46, 2016.

[2] J.-M. Valin et al., "Robust sound source localization using a microphone array on a mobile robot," in *IROS*. IEEE, 2003, vol. 2, pp. 1228–1233.

[3] H. Sawada et al., "Multiple source localization using independent component analysis," in *IEEE Antennas and Propagation Society International Symposium*. IEEE, 2005, vol. 4, pp. 81–84.

[4] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "Tdoa-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1976–1990, 2018.

[5] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," *Sensors*, vol. 18, no. 10, pp. 3418, 2018.

[6] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *ICRA*. IEEE, 2018, pp. 74–79.

[7] S. Adavanne et al., "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[8] K. Shimada et al., "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP*. IEEE, 2021, pp. 915–919.

[9] H. Sundar et al., "Raw waveform based end-to-end deep convolutional network for spatial localization of multiple acoustic sources," in *ICASSP*. IEEE, 2020, pp. 4642–4646.

[10] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," in *MLSP*. IEEE, 2020, pp. 1–6.

[11] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[12] C. Subakan et al., "Attention is all you need in speech separation," in *ICASSP*. IEEE, 2021, pp. 21–25.

[13] E. Tzinis et al., "Improving universal sound separation using sound classification," in *ICASSP*. IEEE, 2020, pp. 96–100.

[14] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[15] J. Le Roux et al., "Sdr–half-baked or well done?," in *ICASSP*. IEEE, 2019, pp. 626–630.

[16] D. Yu et al., "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*. IEEE, 2017, pp. 241–245.

[17] I. Goodfellow et al., "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[18] H. W. Löllmann et al., "The locata challenge data corpus for acoustic source localization and tracking," in *IEEE 10th Sensor Array and Multichannel Signal Processing Workshop*. IEEE, 2018, pp. 410–414.

[19] V. Panayotov et al., "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[20] J. Cosentino et al., "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[21] E. Tzinis et al., "Two-step sound source separation: Training on learned latent targets," in *ICASSP*. IEEE, 2020, pp. 31–35.

[22] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *ICLR*, 2017.

[23] A. M. Torres et al., "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1511–1520, 2012.

[24] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 182–195.