論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Multitask Training of Multi-channel Speaker Separation and Room Acoustic Parameter Estimation
著者(和文)	HartantoRoland, 篠田浩一
Authors(English)	Roland Hartanto, Sakriani Sakti, Koichi Shinoda
出典(和文)	日本音響学会第153回(2025年春季)研究発表会講演論文集, , , pp. 233-234
Citation(English)	, , , pp. 233-234
発行日 / Pub. date	2025, 3

Multitask Training of Multi-channel Speaker Separation and Room Acoustic Parameter Estimation

☆ Roland Hartanto (Science Tokyo), Sakriani Sakti (NAIST),
Koichi Shinoda (Science Tokyo)

1 Introduction

Speaker separation focuses on extracting individual speech signals from a speech mixture. It is applied for single and multi-channel front-end speech processing to deal with overlapping speech. Multi-channel separation leverages spectral and spatial information of speakers, improving separation.

Deep learning methods for multi-channel speech separation have been widely explored. Permutation Invariant Training (PIT) [1] is an approach for training speech separation models, minimizing separation loss across all possible output-target pair permutations. Other studies show that using location information can help improve separation. For example, Location-Based Training (LBT) [2] leverages the direction-of-arrival (DoA) of speakers to organize target speech based on their DoA for loss computation and performs better than PIT. MSDET [3] performs multitask learning of speaker separation and DoA estimation, further improving the separation. However, speaker locations are insufficient to handle various acoustic conditions. In real environments, many parameters can affect acoustic conditions, such as room size, wall surface materials, microphone array locations, and speaker locations.

This work proposes simultaneously learning the speaker separation task with room acoustics parameters estimation, speaker localization, and microphone array localization, exploiting room acoustics information to improve separation in various acoustic conditions. Separation models implicitly learn room acoustics. Multitask learning allows explicit supervision to learn room acoustic parameters, improving separation. Our method separates speech from room acoustic features, capturing reverberation information. Better separation improves the estimation of room acoustic parameters.

2 Previous Studies

Speaker location information, e.g. DoA, can help improve speech separation. Previous work performs multitask speaker separation and DoA estimation training (MSDET) [3] to exploit DoA information. Multitask learning [4] enhances the performance of individual tasks, leveraging information from multiple tasks to learn a shared representation. Since speech signals are affected by many room acoustics parameters, DoAs alone cannot handle diverse conditions.

Some studies focus on estimating room acoustics parameters such as the volume [5, 6], geometry [7], and total surface area of the room [6], the average absorption coefficient of room surfaces [6, 7], and reverberation time [7]. Another study [8] jointly estimates reverberation time and performs speech dereverberation to help improve speech dereverberation. Although this method is effective, estimating reverberation time alone may not be sufficient.

3 Proposed Method

We train a speaker separation model by performing multitask learning of separation (SS), speaker localization (SL), microphone array localization (ML), and room parameters estimation (RP). We utilize SpatialNet [9] as our speaker separator backbone. SpatialNet performs an attention mechanism on each narrow frequency band that contains spatial information of each speaker, which is crucial for multi-channel speaker separation.

Fig. 1 illustrates our proposed system. Unlike [9], this architecture has a unit for SS and SL for each speaker, named speaker output unit, and a unit for ML and RP, named room acoustics unit. The speaker output unit consists of a linear layer to output speech spectrum and features for SL and a linear layer to output speaker location in (x, y). The room acoustics unit consists of a linear layer to output features for ML and RP, a linear layer to output microphone array location in (x, y), and a linear layer to output room parameters. The room parameters include reverberation time, early decay time, volume, surface area, width, length, average absorption coefficient of room surfaces, direct-to-reverberant ratio,

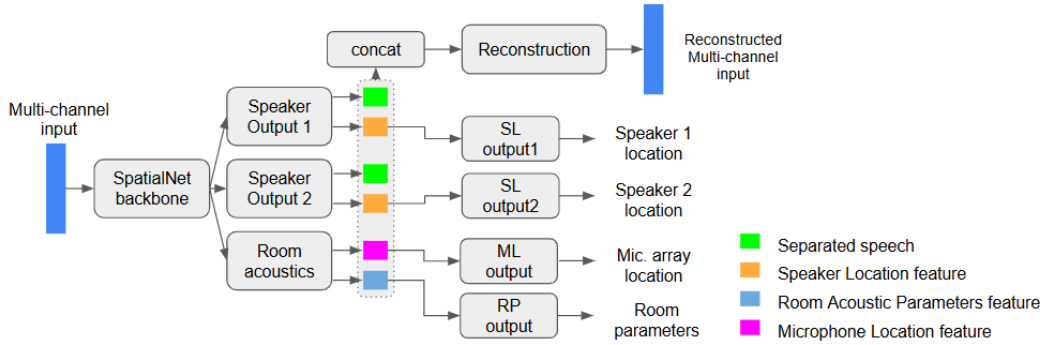

Fig. 1 Model Architecture

Table 1 Evaluation results on SMS-WSJ-Plus

System	SI-SDR (dB)	WER (%)
SpatialNet+PIT [9]	**13.8**	15.89
SpatialNet+MSDET	12.8	16.56
SpatialNet+Proposed	13.4	**15.22**

Table 2 Separation performance on CHiME 6

System	WER (%)
SpatialNet+PIT	59.73
SpatialNet+Proposed	**57.09**

and clarity. We also add a convolutional layer to reconstruct the input mixture using the separated speech, the features of SL, ML, and RP.

The model is trained in several steps. The first step trains each task iteratively. We train SS and SL and freeze the output layer parameters of ML and RP for j epochs. Next, we train ML, freeze the parameters of SS, SL, and RP, and then similarly train RP. After that, we repeat the process from training SS-SL, ML, to RP for k iterations. The second step trains two tasks simultaneously for each iteration. The training starts from SS-SL to ML-RP for l iterations. Finally, all tasks are trained simultaneously until convergence.

The multitask loss is the weighted sum of the SS, SL, ML, and RP losses. The SS loss is the negative Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) loss between the ground truth and the separated speech in the time domain. The SL, ML, RP, and reconstruction losses are the mean squared error (MSE) loss. The multitask loss can be expressed as follows:

$$\begin{aligned} L = &(1 - w_{\text{ML}} - w_{\text{RP}} - w_{\text{recons}}) \\ &((1 - w_{\text{SL}})L_{\text{SS}} + w_{\text{SL}}L_{\text{SL}}) + \\ &w_{\text{ML}}L_{\text{ML}} + w_{\text{RP}}L_{\text{RP}} + w_{\text{recons}}L_{\text{recons}}. \end{aligned} \quad (1)$$

4 Experiments

We use SMS-WSJ-Plus [9], a simulated noisy and reverberant speaker separation dataset containing two-speaker speech mixtures. It simulates a circular array with six microphones. The reverberation time range is [0.1, 1.0] s. The room length and width range are [4, 10] m. The room height range is [3, 4] m. We present the experiment results in Tab. 1. We use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) and Word Error Rate (WER) to evaluate separation performance. Our method (SpatialNet+Proposed) performs worse (0.4 points) in SI-SDR than the one trained using only PIT (SpatialNet+PIT). However, it achieves a lower WER by 0.67 points. Our proposed method also performs better than the one trained using MSDET.

We also evaluate our method on CHiME 6, a real environment dataset recorded using six units of 4-channel mic arrays. The results are shown on Tab. 2. Our method outperforms SpatialNet+PIT.

5 Conclusion

The multitask approach of separation and room acoustic parameters estimation improves separation, especially regarding WER. It exploits various room acoustic parameters to improve speaker separation.

Reference

[1] Kolbæk *et al.*, TASLP, vol.25, 1901 - 1913, 2017
[2] Taherian *et al.*, TASLP, vol.30, 2791 - 2800, 2022
[3] Hartanto *et al.*, Interspeech, 2170-2174, 2024
[4] Caruana, Machine Learning, vol.28, 41–75, 1997.
[5] Genovese *et al.*, ICASSP, 231-235, 2019
[6] Srivastava *et al.*, WASPAA, 226-230, 2021
[7] Yu, Kleijn, TASLP, vol.29, 436 - 447, 2021
[8] Wu *et al.*, EURASIP, 81, 2017
[9] Quan *et al.*, TASLP, vol.32, 1310 - 1323, 2024