論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Multitask Learning of Speaker Separation and Direction-of-Arrival Estimation
著者(和文)	Hartanto Roland, 篠田 浩一
Authors(English)	Roland Hartanto, Sakriani Sakti, Koichi Shinoda
出典(和文)	日本音響学会第151回(2024年春季)研究発表会 講演論文集, , , pp. 69-70
Citation(English)	, , , pp. 69-70
発行日 / Pub. date	2024, 3

Multitask Learning of Speaker Separation and Direction-of-Arrival Estimation

☆ Roland Hartanto (Tokyo Tech), Sakriani Sakti (JAIST), Koichi Shinoda (Tokyo Tech)

1 Introduction

Speech separation is the process of separating individual speaker voices from a mixture of multiple speakers' voices. Speech separation techniques have been developed for monaural and multichannel speech processing. Multichannel separation utilizes spectral and spatial information of speech sources, which help improve separation performance.

Deep learning-based speech separation techniques have been extensively studied. Permutation Invariant Training (PIT) [1] is commonly used in speech separation model training. It trains the model by minimizing separation loss over all possible output-target permutations. However, this technique is costly as the number of speakers increases. A previous work called Location-Based Training (LBT) [2] attempted to utilize the direction-of-arrival (DOA) of speakers to support separation model training. It solves the permutation problem by ordering the target speech according to their DOA for loss calculation and performs better than PIT. However, LBT does not consider the cycle of DOA, which may cause confusion when assigning separation outputs because a source located between 0-90 degrees is considered distant from one located between 270-360 degrees.

Our work explores the use of sound sources' DOA to improve speaker separation. To solve the aforementioned problems, we employ multitask learning of speaker separation and DOA estimation. The DOA information of each speaker is explicitly used in the multitask loss calculation as supervision in addition to the target speech.

2 Previous Studies

In multichannel speech separation, DOA information is useful to help improve speech separation performance. [Z. Chen et al., 2018] uses the DOA label of the target speaker to perform the separation during inference. However, DOA labels are not available during the inference stage. [W.Sun et al., 2022] jointly trains the separator and DOA estimator in a multitask manner for a single target speaker. It improves target speaker separation performance as multitask learning [5] allows learning a shared representation by using information from different tasks and improves the performance of each task. Nevertheless, this approach cannot be applied directly to separate multiple speakers as it needs the target speaker's information.

[H.Taherian et al., 2022] introduced an approach called Location-Based Training (LBT). It solves the permutation problem of speaker separation by using DOA labels to order the source targets for loss calculation in the training stage. However, LBT does not consider the DOA cycle. Confusion may occur in separation since a source located between 0-90 degrees is treated far away from one located between 270-360 degrees.

3 Proposed Method

We perform multitask learning by explicitly pairing each source with its corresponding DOA in our multitask loss calculation. Our study uses TFGridNet [6] as our speech separator baseline, which directly outputs the separated signals for all speakers. We chose it because it separates real and imaginary speech frequency components and uses an attention mechanism to achieve strong temporal modeling for the whole utterance.

We illustrate our system when the number of speakers is two in Fig. 1. We add DOA estimation layers after the separation layer in TFGridNet to perform DOA estimation. The estimated DOAs for output Speech 1 and Speech 2 are DOA 1 and DOA 2, respectively. We estimate DOA by using a classification approach. The final feed-forward layer output size is the number of DOA classes determined by $(360/r)$, where r is the DOA resolution.

We compute the multitask loss by performing the weighted sum of the separation and DOA estimation losses. The separation loss is the L1 loss between the ground truth and the generated signals in the time-frequency domains [6]. The DOA estimation loss is the cross-entropy loss between the predicted and the

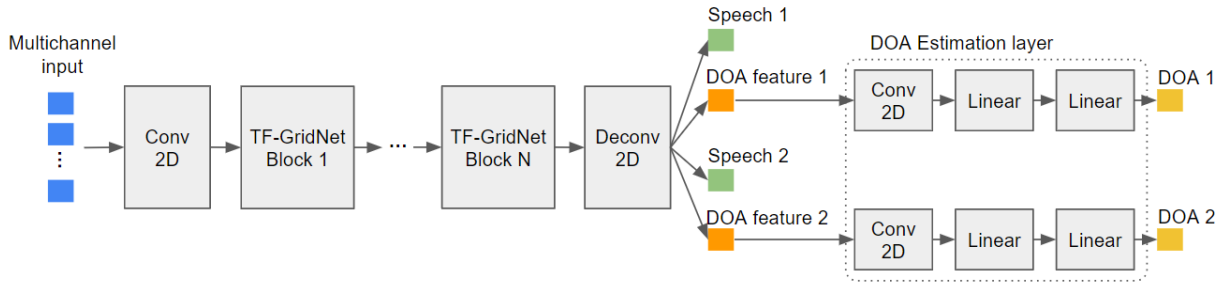

Fig. 1 Multitask Speech Separator and DOA estimator

Table 1 Speech separation performance.

Scenarios	SI-SDR (dB)
TF-GridNet [6]	19.90
LBT [2]	13.22
Multitask	**20.21**

Table 2 DOA estimation performance

Scenarios	MAE (degrees)
Baseline DOA est.	0.64
Multitask	**0.52**

ground truth DOAs [7]. The multitask loss can be written as follows:

$$L_{\text{Multitask}} = (1 - w_{\text{DOA}})L_{\text{sep}} + w_{\text{DOA}}L_{\text{DOA}}. \quad (1)$$

4 Experiments

We use SMS-WSJ [8], a simulated dataset with reverberation for multichannel source separation derived from the WSJ corpus. The dataset contains two-speaker speech mixtures. It simulates a circular microphone array with six microphones.

We present our experiment results in Tab. 1 and Tab. 2. We use the Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) to evaluate speech separation performance and Mean Absolute Error (MAE) for DOA estimation. The LBT in Tab. 1 uses Dense-UNet architecture [2]. Our separator and DOA estimator baselines are TFGridNet trained using only separation loss and DOA estimation loss, respectively. The DOA resolution applied in the experiments is 1 degree. Our method outperforms the baseline separator and LBT for speech separation. In addition, our method also exhibits better DOA estimation performance than the baseline DOA estimator.

5 Conclusion

The multitask approach of separation and DOA estimation improves the separation and DOA estimation performances. We explicitly pair each source to its DOA in our multitask loss, allowing the model to separate sources originating from different directions better than LBT.

Reference

[1] M. Kolbaek et al., "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," TASLP, vol. 25, no. 10, 2017

[2] H. Taherian et al., "Multi-channel talker-independent speaker separation through location-based training," TASLP, vol. 30, pp. 2791–2800, 2022

[3] Z. Chen et al., "Multi-channel overlapped speech recognition with location guided speech extraction network," in IEEE SLT, pp. 558–565, 2018

[4] W. Sun et al., "Spatial aware multi-task learning based speech separation," arXiv:2207.10229, 2022.

[5] R. Caruana, "Multitask learning," Machine Learning 28, 41–75, 1997.

[6] Z.Q. Wang et al., "TF-GridNet: Integrating full- and sub-band modeling for speech separation," TASLP, vol. 31, pp. 3221-3236, 2023.

[7] A.S. Subramanian et al., "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition," CSL, vol. 75, pp. 101360, 2022

[8] L. Drude et al., "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," arXiv:1910.13934, 2019