Speaker-Aware Mixture of Mixtures Training for Weakly Supervised Speaker Extraction

*Zifeng Zhao, Rongzhi Gu, Dongchao Yang, Jinchuan Tian, Yuexian Zou**

ADSPLAB, School of ECE, Peking University, Shenzhen, China

{zhaozifeng, dongchao98, tianjinchuan}@stu.pku.edu.cn, {1701111335, zouyx}@pku.edu.cn

Abstract

Dominant researches adopt supervised training for speaker extraction, while the scarcity of ideally clean corpus and channel mismatch problem are rarely considered. To this end, we propose speaker-aware mixture of mixtures training (SAMoM), utilizing the consistency of speaker identity among target source, enrollment utterance and target estimate to weakly supervise the training of a deep speaker extractor. In SAMoM, the input is constructed by mixing up different speaker-aware mixtures (SAMs), each contains multiple speakers with their identities known and enrollment utterances available. Informed by enrollment utterances, target speech is extracted from the input one by one, such that the estimated targets can approximate the original SAMs after a remix in accordance with the identity consistency. Moreover, using SAMoM in a semi-supervised setting with a certain amount of clean sources enables application in noisy scenarios. Extensive experiments on Libri2Mix show that the proposed method achieves promising results without access to any clean sources (11.06 dB SI-SDRi)[1]. With a domain adaptation, our approach even outperformed supervised framework in a cross-domain evaluation on AISHELL-1.

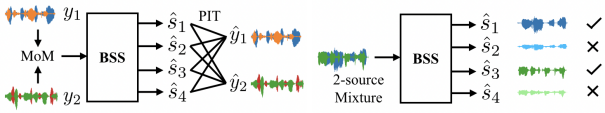
Index Terms: speech separation, target speaker extraction, weakly supervised learning, domain adaptation

1. Introduction

Speech separation is a fundamental component in many speech processing systems, for example, acting as a front-end module for robust automatic speech recognition (ASR). Without such a front-end, the performances of downstream tasks may deteriorate greatly, especially when an interfering speaker exists.

Over the decades, lots of efforts have been made to crack this problem. One direction is to extract target speech with the auxiliary of an enrollment utterance from the target speaker. Following supervised learning paradigm [1], dominant researches formulate target speaker extraction (TSE) as a supervised learning problem, based on which various deep models were proposed to advance the best performance [2][3][4]. In such a framework, artificially generated multi-speaker mixtures and corresponding clean sources are given as sample pairs for training. Informed by an additional enrollment utterance, a deep model consumes the input and extracts the target out of the mixture, such that the output estimate approximates target speaker's speech.

Such a mix-and-separate paradigm, however, has two major drawbacks. First, corpus with adequate clean utterances is required, serving as training ground truth as well as for simulating input mixtures. Second, even if abundant simulated data is available, the model's performance in real-world scenario can still be poor, since there is usually a channel mismatch between the simulated data and the target domain.



(a) *During training* (b) *During inference*

Figure 1: *MixIT for SS models. A mismatch of the source number between (a) training and (b) inference may result in over-separaion problem.*

A feasible solution to these problems is to ameliorate current fully supervised framework with unsupervised or weakly supervised learning. Wisdom et al. proposed mixture-invariant training (MixIT) for speech separation (SS) [5][6], where the model consumes a mixture of mixtures, and estimates multiple sources as outputs by one pass, one output channel for each latent source. In this method, a mismatch of output channel number may exist between training and inference, and hence lead to suboptimal performance due to over-separation [7], which is illustrated in Figure 1. Very recent works also explored utilizing weak speaker labels to train TSE models, especially by making use of a pretrained speaker encoder [8][9] to form a speaker identity loss. These approaches, however, require an additional pretrained model, and may result in under-separation due to the robustness of the extra speaker encoder.

In this paper, we propose speaker-aware mixture of mixtures training (SAMoM), by making advantages of the speaker identity consistency among target source, enrollment utterance and target estimate, to weakly supervise the training of a speaker extraction model. Without access to clean sources, the input is constructed by mixing up different speaker-aware mixtures (SAMs), each contains multiple speakers with their identities known and corresponding enrollment utterances available. Informed by enrollment utterances, target speech is extracted from the input one by one, such that the estimated targets can approximate the original SAMs after a remix in accordance with the identity consistency. The proposed method is feasible since speaker identity labels and enrollment utterances are usually much easier to obtain than abundant clean speech from the target domain. Moreover, when a certain amount of clean speech is available, SAMoM can be further extended for noisy scenarios by substituting one of the SAM with a single-speaker clean speech, while the other ingredient of the input is still a SAM, but a noisy recording with ambient sound. The input mixture is then the sum of a clean speech and a noisy SAM, while other parts of the training remains the same as the naïve SAMoM framework, forming a semi-supervised training paradigm.

The rest of paper is organized as follows. Section 2 reviews

* Corresponding author.

[1] Some audio samples of the model's output are available at our page: https://zhazhafon.github.io/demo-samom/

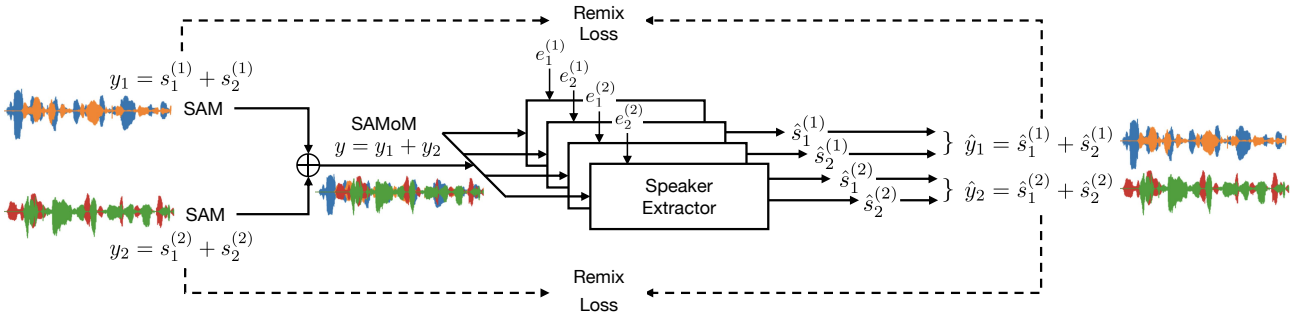


Figure 2: *The proposed SAMoM training framework.*

speaker extraction. Section 3 presents the proposed framework and its noisy extension. Experiments and detailed results are reported in Section 4. We conclude this paper in Section 5.

2. Target speaker extraction

2.1. Problem formulation

In near-field speech separation, the speech mixture can be considered as a linear combination of speakers' speech and ambient noise:

$$y = \sum_{j=1}^{J} s_j + n \tag{1}$$

where J is the number of existing speakers, s_j for $j = 1, \ldots, J$ is the speech signal of the jth speaker, n is the additive noise and y is the speech mixture. Speaker extraction is essentially a guided speech separation, where the auxiliary information is an enrollment utterance of the target speaker:

$$\hat{s}_t = SpkExtr(y|e_t;\theta) \tag{2}$$

where e_t is the enrollment utterance, $SpkExtr()$ denotes a TSE model with parameter θ and the output $\hat{s}_t$ is the target speech estimate. Following supervised separation [1], mainstream researches treat speech extraction as a supervised learning problem, in which abundant clean sources are given as ground-truth targets. Speech mixtures are artificially generated by mixing up these clean sources according to Eq. 1. Then mixtures and clean sources are used as sample pairs for supervised training, acting as model's inputs and labels respectively.

3. Methods

3.1. Mixture of mixtures

Supervised speech separation models are trained on mixture of sources (MoS), in which model's input is generated by mixing up different clean sources. Differently, the mixture of mixtures (MoM) paradigm is to construct an input mixture by mixing up different speech mixtures, and train the model to reconstruct those original mixtures by a properly designed loss function, such that the model acquires an ability to separate a mixture of sources for inference. Some previous researches adopted MoM as their model's input, especially in audio-visual speech separation [10] and unsupervised speech separation [5][6][7], proving the effectiveness of such a paradigm.

3.2. Speaker-aware mixture of mixtures training

To eliminate the over-separation problem introduced in Section 1, we make use of auxiliary speaker information and identity consistency. As illustrated in Figure 2, the proposed speaker-aware mixture of mixtures training (SAMoM) framework can be divided into three phases: mixture generation for input audio creation, speaker extraction for target source estimation and SAM remix for remixing estimates to calculate a signal-level loss function. The overall framework is designed in a weakly supervised setup and has no access to any clean sources.

Mixture Generation. Speaker-aware mixture (SAM) is used as a basic material for training in the proposed framework. Generally, a SAM is a mixture consisting of speech from multiple speakers, with their identities known and some enrollment utterances available, both of which are utilized as weak labels during training. The enrollment utterance provides target-related clue for speaker extraction, while the speaker identity is used to guide the subsequent remix process. Note that in naïve SAMoM, we assume no noise interference and thus the SAM is a linear combination of speech from different speakers:

$$y_i = \sum_{j=1}^{J_i} s_j^{(i)} \tag{3}$$

where $s_j^{(i)}$ for $j = 1, \ldots, J_i$ is the speech signal from the jth speaker in the ith SAM. J_i is the total speaker number in the ith SAM and y_i denotes the ith SAM. Different SAMs should not have any speakers in common, and this can be easily realized by checking their speaker labels. With multiple SAMs available, the input is generated by:

$$y = \sum_{i=1}^{N} y_i \tag{4}$$

where N is the number of SAMs, y is the input audio to the TSE model. An example of two SAMs each containing two sources ($N = 2, J_1 = J_2 = 2$) is depicted in Figure 2.

Target Speaker Extraction. By informing the model of enrollment utterances that belonging to different speakers, the corresponding target speech is extracted from the input mixture one by one following Eq. 2:

$$\hat{s}_j^{(i)} = SpkExtr(y|e_j^{(i)};\theta), \forall i, \forall j \tag{5}$$

where $\hat{s}_j^{(i)}$ and $e_j^{(i)}$ are the speech estimate and enrollment utterance for the jth speaker in the ith SAM. $e_j^{(i)}$ can be of any length without corresponding to y, while long recordings generally bring about better results. Technically, this process can be done either in sequence or in parallel for different speakers, as long as the extraction for them is uncorrelated. Besides, correlated extraction methods (e.g. recursive separation [11]) can also be employed, but we leave this to future research.

SAM Remix. According to Sec. 2, for a certain speaker k, the target source estimate $\hat{s}_k$ is extracted from a SAM containing the source s_k, with speaker's enrollment utterance e_k. There is always a speaker identity consistency among these three signals: the enrollment utterance e_k, the target source estimate $\hat{s}_k$ and the target source s_k. In the last stage, estimated sources are remixed in accordance with such a consistency, so that the remixed mixtures can approximate the original SAMs:

$$\hat{y}_i = \sum_{j=1}^{J_i} \hat{s}_j^{(i)} \tag{6}$$

where notations are consistent with Eq. 3. Take Figure 2 as an example, $\hat{s}_1^{(1)}$ and $\hat{s}_2^{(1)}$ are extracted fom y_1 for speaker 1 and speaker 2 with their enrollment utterances $e_1^{(1)}$ and $e_2^{(1)}$. According to the identity consistency, $\hat{s}_1^{(1)}$ and $\hat{s}_2^{(1)}$ are remixed to form $\hat{y}_1$ so as to reconstruct y_1. Finally, the remix loss is formed by applying a negative scale-invariant signal-to-distortion ratio (SI-SDR) [12] between the original SAMs and the remixed SAMs, similar to the co-separation loss proposed in [10]:

$$L = \frac{1}{N}\sum_{i}^{N} L_i = \frac{1}{N}\sum_{i}^{N} l(y_i, \hat{y}_i) \tag{7}$$

$$l(x, \hat{x}) = -10log_{10}(\frac{||\alpha x||^2}{||\alpha x - \hat{x}||^2}), \alpha = \frac{< x, \hat{x} >}{||x||^2} \tag{8}$$

Domain Adaptation. Furthermore, since SAMoM does not require any clean sources for training, it can adapt to the testing data through an additional fine-tuning. With weak speaker labels available, this can be done by training the model on the testing data for a certain epochs with low learning rate, after which the model learns the channel characteristics of the testing data and may generalize better. Such an adaptation capability can be extremely helpful when there is a channel mismatch between training and testing domain.

3.3. Extension to noisy scenario

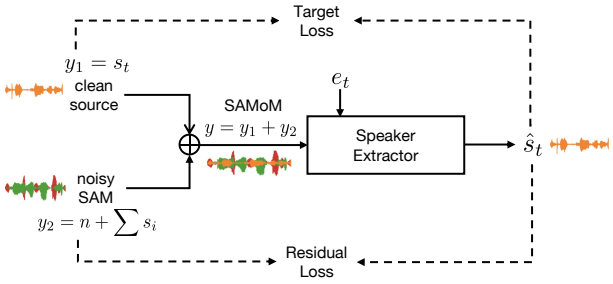


Figure 3: *Extension to noisy scenario.*

The proposed SAMoM framework can be extended to a noisy setup for more general applications, but this may require a certain amount of single-speaker utterances as clean ground truths, which violates the weak supervision setting and results in a semi-supervised training paradigm. As depicted in Figure 3, one of the SAM is substituted by a clean speech with a single speaker present, while the other ingredient of the input is still a SAM, but a noisy recording with ambient sound. The only target speaker is set to be that of the clean source and the input is the sum of a clean speech and a noisy mixture. A source estimate is extracted from the input given the enrollment utterance of the target speaker. The loss function is composed of a target loss and a residual loss:

$$L = \frac{1}{2}(L_{target} + L_{residual}) \tag{9}$$

$$L_{target} = l(s_t, \hat{s}_t) \tag{10}$$

$$L_{residual} = l(y_2, y - \hat{s}_t) \tag{11}$$

where s_t is the clean source, $\hat{s}_t$ is the target source estimate, $y - \hat{s}_t$ is the estimation for the residual signal.

4. Experiments

4.1. Datasets

The proposed framework is evaluated under three different tasks and all audios are downsampled to 8kHz in advance.

In the first task, the proposed framework is trained with only weak speaker labels and compared with fully supervised methods. We use Libri2Mix dataset [13] for this task. *train-100* is used as our training set, and *dev* and *test* subsets for validation and evaluation. Different from speech separation, speaker extraction requires additional enrollment utterances. To this end, we utilize an enrollment list[2] for Libri2Mix. Note that speech mixtures with three or even more sources (e.g. Libri3Mix) are also compatible with the proposed framework.

The second task is a cross-domain evaluation. We created a dataset with channel characteristics different from the first task. The proposed dataset is referred to as *aishell1-2mix*, which is simulated based on AISHELL-1 [14]. Test audios are directly generated by mixing up 2 randomly selected utterances of different speakers, without scaling. Note that more complicated mixing tricks can be used (e.g.random SNR sampling [15] or loudness control [13]). Both *dev* and *test* subsets of AISHELL-1 are used to construct our evaluation set *eval*, while the training set is not used. More details of the proposed *aishell1-2mix* are listed in Table 1.

In the last task, the noisy extension of the proposed framework is evaluated. Noise from WHAM! dataset [16] is used together with speech from task 1 to simulate noisy mixtures.

	Libri2Mix / test set	***aishell1-2mix* / eval set**
#Speakers	40	60
#Utterances	3000	2500
Hours	11	2.08
Language	English	Chinese

Table 1: *A comparison between the test set of Libri2Mix and the evaluation set of aishell1-2mix.*

4.2. Network configuration

TD-SpeakerBeam [17] is employed for our experiments as TSE model, which combines the speaker clue fusion mechanism of previous works on SpeakerBeam [18] and the time-domain convolutional separation network in Conv-TasNet [19]. Note that proposed training methods can be applied to any off-the-shelf TSE models, but we chose TD-SpeakerBeam such that we can fairly compare our results with those by a similar speech separation network (Conv-TasNet) using MixIT. Conv-TasNet is used for both supervised SS and unsupervised MixIT baselines, and the hyperparameters and network architecture are set identical

[2]https://github.com/BUTSpeechFIT/speakerbeam/tree/main/egs/libri2mix

	SI-SDRi (dB)	SDRi (dB)	STOI	PESQ
sup SS	**13.40**	**13.82**	**0.92**	2.74
sup TSE	12.86	13.40	0.90	**2.75**
unsup MixIT	5.72	6.92	0.79	1.98
SAMoM	8.97	9.80	0.85	2.28
+Adaptation	**11.06**	**11.64**	**0.88**	**2.41**

Table 2: *Performance of different training methods for speech separation and speaker extraction on Libri2Mix.*

with the separator of TD-SpeakerBeam. All models are implemented using the Asteroid toolkits [20] and trained for 100 epochs with learning rate of $1e^{-3}$. During training, both input mixtures and enrollment speech are 3-second audio segments randomly cut from the original utterances. While for inference, full-length utterances are used.

4.3. Results

For a more complete evaluation, we compare different methods on four metrics[3]: two signal-level metrics (SI-SDRi and SDRi), one speech intelligibility metric (STOI) and one speech quality metric (PESQ).

Proposed method and baselines. In the first task, we compared our method with several baselines on Libri2Mix: (1) *sup SS*: speech separation with supervised training, (2) *sup TSE*: speaker extraction with supervised training and (3) *unsup MixIT*: speech separation with MixIT unsupervised training. Permutation-invariant training [21][22] is adopted for all speech separation models (*sup SS* and *unsup MixIT)* during both training and inference. While for speaker extraction models, the enrollment list introduced in Section 4.1 is used. Results are reported in Table 2. As illustrated in the first two rows, *sup TSE* is slightly inferior to *sup SS*, probably due to that speaker extraction comes across with speaker bias in some enrollment utterances; MixIT [5] with purely unsupervised training (*unsup MixIT*) achieved a SI-SDRi of 5.72 dB, which is far less than the fully supervised speech separation baseline (13.40 dB SI-SDRi). This is because a mismatch of output channel number exists between training and inference, which leads to over-separation during testing, as depicted in Figure 1. Our proposed method (*SAMoM*) significantly outperforms *unsup MixIT* by more than 3 dB in terms of SI-SDRi. Furthermore, we performed a domain adaptation as introduced in Section 3.2, which is essentially a fine-tuning with learning rate of $1e^{-4}$ for 20 more epochs on the test set using weak speaker labels. *SAMoM+Adaptation* achieved a SI-SDRi of 11.06 dB, which is pretty close to that of *sup TSE* (12.86 dB SI-SDRi). A sample data of *SAMoM+Adaptation* is depicted in Figure 4.

	SI-SDRi (dB)	SDRi (dB)	STOI	PESQ
sup TSE	1.99	2.65	0.68	1.77
+Adaptation	4.56	5.48	0.73	2.06
SAMoM	0.73	1.97	0.66	1.72
+Adaptation	**5.86**	**6.64**	**0.75**	**2.12**

Table 3: *Cross-domain evaluation on aishell1-2mix.*

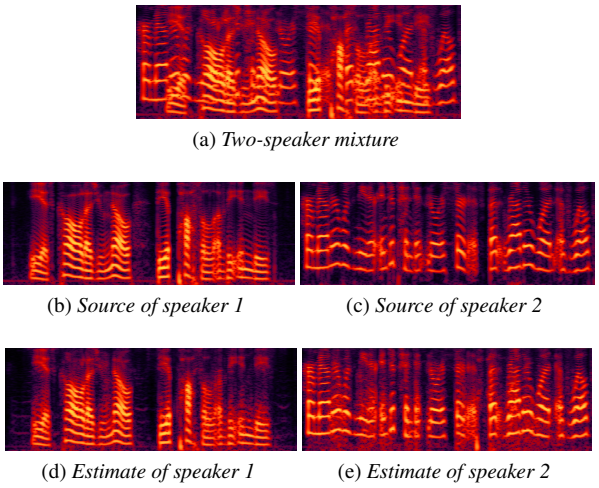
Cross domain evaluation. The second task is to show model's generalization ability when applied to a new scenario with completely different channel characteristics. Two base models were used in this task: (1) *sup TSE*: fully supervised training, and (2) *SAMoM*: weakly supervised training. Both of the models were trained on Libri2Mix with the same setups as task 1, but evaluated on the proposed *aishell1-2mix*. As illustrated in Table 3, although *sup TSE* performs better than *SAMoM*, both of their performance are very poor when confronted with a different channel characteristic. To ease such a channel mismatch, a domain adaptation can be done by fine-tuning base models in the target domain with weak labels according to Secion 3.2 (*+Adaptation*). To this end, base models are fine-tuned on *aishell1-2mix* for 20 epochs, with an initial learning rate of $1e^{-3}$ and halved at the antepenultimate epoch. With domain adaptation, the performance of *sup TSE* and *SAMoM* increased to 4.56 dB and 5.86 dB SI-SDRi, respectively, showing that such a domain adaptation can play a crucial role for cross-domain inference.

[3]https://github.com/fgnt/pb_bss

	SI-SDRi (dB)	SDRi (dB)	STOI	PESQ
Supervised	**10.79**	**11.51**	**0.83**	**2.15**
Proposed	9.55	10.26	0.81	1.99

Table 4: *Performance under noisy condition.*

Noisy extension. As illustrated in Table 4, SAMoM's semi-supervised extension achieved a SI-SDRi of 9.55 dB, which is close to the fully supervised model (10.79 dB). This suggests the effectiveness of the proposed method.



(a) *Two-speaker mixture*

(b) *Source of speaker 1* (c) *Source of speaker 2*

(d) *Estimate of speaker 1* (e) *Estimate of speaker 2*

Figure 4: *Spectrograms of (a) two-speaker mixture, (b)(c) target source signals and (d)(e) estimated signals on Libri2Mix.*

5. Conclusions

In this paper, we propose speaker-aware mixture of mixtures training (SAMoM), a weakly supervised learning framework for speaker extraction. The proposed method achieves considerable results with only weak speaker labels accessible. Since no clean sources are required for training, it can realize domain adaptation to reduce performance attenuation caused by channel mismatch. In addition, we extend it for noisy condition with semi-supervised learning. Extensive experiments on LibriMix and AISHELL-1 validate the effectiveness of our methods.

6. Acknowledgements

This paper was partially supported by the Shenzhen Science and Technology Fundamental Research Program (No:GXWD20201231165807007-20200814115301001) and Natural Science Foundation of China (NSFC 62176008).

7. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5554–5558.

[3] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[4] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1370–1384, 2020.

[5] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[6] A. Sivaraman, S. Wisdom, H. Erdogan, and J. R. Hershey, "Adapting speech separation to real-world meetings using mixture invariant training," *arXiv preprint arXiv:2110.10739*, 2021.

[7] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "Teacher-Student MixIT for Unsupervised and Semi-Supervised Speech Separation," in *Proc. Interspeech*, 2021, pp. 3495–3499.

[8] K. Zmolikova, M. Delcroix, D. Raj, S. Watanabe, and J. Černockỳ, "Auxiliary loss function for target speech extraction and recognition with weak supervision based on speaker characteristics," *Proc. Interspeech*, pp. 1464–1468, 2021.

[9] F. Pishdadian, G. Wichern, and J. Le Roux, "Learning to separate sounds from weakly labeled scenes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 91–95.

[10] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3879–3888.

[11] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive Speech Separation for Unknown Number of Speakers," in *Proc. Interspeech*, 2019, pp. 1348–1352.

[12] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[13] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[14] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[15] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[16] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Proc. Interspeech*, 2019, pp. 1368–1372.

[17] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 691–695.

[18] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6965–6969.

[19] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[20] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[21] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[22] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.