SpkAugTSE: A Simple and Efficient Approach to Address Target Confusion in End-to-End Speaker Extraction

Zhenghai You* and Zhenyu Zhou* and Lantian Li* and Dong Wang†
* School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
E-mail: lilt@bupt.edu.cn
† Center for Speech and Language Technologies, BNRist, Tsinghua University, China
E-mail: wangdong99@mails.tsinghua.edu.cn

***Abstract*—Target confusion, defined as occasional switching to non-target speakers, poses a key challenge for end-to-end speaker extraction (E2E-SE) systems. We argue that this problem is largely caused by the lack of generalizability and discrimination of the speaker embeddings, and introduce a simple yet effective speaker augmentation strategy to tackle the problem. Specifically, we propose a time-domain resampling and rescaling pipeline that alters speaker traits while preserving other speech properties. This generates a variety of pseudo-speakers to help establish a generalizable speaker embedding space, while the speaker-trait-specific augmentation creates hard samples that force the model to focus on genuine speaker characteristics. Experiments on WSJ0-2Mix and LibriMix show that our method mitigates the target confusion and improves extraction performance. Moreover, it can be combined with metric learning, another effective approach to address target confusion, leading to further gains.**

I. Introduction

Extracting a specific speaker's voice from multi-talker speech signals is a fundamental challenge in speech signal processing, commonly referred to as **S**peaker **E**xtraction (SE) [1]–[3]. Unlike Speaker Separation (SS) [4]–[6], which aims to separate all speakers in a mixed speech, SE refers to an enrollment utterance of the target speaker and selectively extracts the speech of the target speaker.

In recent years, the *end-to-end* speaker extraction (E2E-SE) approach has gained popularity. In this approach, a speaker encoder and a speech encoder are integrated with a speaker extractor, forming a unified end-to-end extraction framework [1], [7]–[9]. Typically, the speech encoder produces latent representations of the mixed speech, while the speaker encoder generates a target speaker embedding of the reference speech. The speaker extractor then extracts the target speaker's speech from the mixed speech representations by referring to the target speaker embedding.

Despite significant advancements, the *target confusion* problem remains a challenging issue [8], [10], [11]. This problem occurs when the system extracts speech components of the interfering speaker as the target speech, resulting in a low-quality listening experience. This problem is particularly severe when the target and interfering speakers share similar vocal characteristics, making them difficult to distinguish. It also occurs when the mixed speech is severely corrupted by noise, in which case the speaker traits of the target and interfering speech are equally difficult to identify, leading to potential target confusion.

Recently, several approaches have been proposed to address the target confusion problem. From the data perspective, Li *et al.* [12] introduced noise and reverberation into enrollment utterances during model training, to improve the robustness of the speaker encoder in complex conditions. From the model perspective, architectures such as SpEx++ [13] and DPRNN-IRA [14] incorporate an iterative refinement mechanism that incrementally enhances the embeddings of the target speaker. From the objective perspective, Zhao *et al.* [10] proposed a metric learning approach, which introduces an additional triplet loss to contrast the distance between the enrollment-target pairs and the enrollment-interfering pairs.

Speaker augmentation has been shown to be very effective in speaker recognition tasks [15]–[17]. Briefly, this approach produces some 'pseudo speakers' by perturbing the vocal fold or vocal tract properties of some existing real speech utterances, or utilizing multi-speaker text-to-speech or voice conversion techniques. This paper explores the possibility of addressing the target confusion problem by speaker augmentation. We choose a simple time-domain resampling and rescaling pipeline. By this approach, we can modify the fundamental frequency (F0) and formants of the original speech, and this modified speech can be regarded as from 'new' speakers or pseudo speakers. Adding the generated speech into the training dataset is supposed to improve the construction of the speaker space, leading to increased generalizability of the speaker embeddings. Most importantly, the augmentation changes the speaker traits only, while preserving all the non-speaker properties, e.g., textual content, tempo, and prosody of the original speech. We argue that when the augmented version and the original version of the same speech are mixed, it forms a "hard" sample for the model training, as it is more difficult to process than usual mixture samples where the target and interfering speech usually possess different content, tempo, and prosody. Our hypothesis is that involving such hard samples in model training will compel the system to capture the genuine

979-8-3315-7206-8/25/$31.00

speaker characteristics, by which the target confusion problem can be mitigated.

We conduct experiments on two benchmark datasets, WSJ0-2Mix [18] and LibriMix [19], demonstrating that our method consistently improves performance with two state-of-the-art E2E-SE architectures, including DPRNN [20] and SpEx+ [21]. The ablation study shows that the performance improvement can be largely attributed to the hard samples produced by speaker augmentation. Additionally, we show that our method can be combined with other advanced techniques, such as metric learning, resulting in further performance improvement.

The rest of this paper is structured as follows. Section II reviews related work and Section III details the proposed speaker augmentation approach. Section IV presents the experimental results, and Section V concludes the paper.

II. Related Work

Speaker augmentation techniques have been widely adopted in speaker recognition tasks [15], [16], [22]. Among these, speed perturbation is perhaps the most popular. It modifies the pitch and formants of a speech utterance while keeping the linguistic content unchanged. This simple yet effective strategy has been shown to be highly effective in speaker recognition under complex conditions [17].

Our work is built upon the speed perturbation approach, though we made a slight change to meet our goal of tackling the key problem of target confusion in target speaker extraction. Specifically, we designed a time-domain resampling and rescaling pipeline that produces speech of pseudo-speakers that differs from the original speech only in speaker traits. This allows us to construct hard samples that can be used to enforce the model to identify genuine speaker traits.

In speaker extraction, Li *et al.* [12] investigated data augmentation on enrollment speech by adding noise and reverberation. Our approach is speaker augmentation, which adds pseudo-speakers. Moreover, we emphasize constructing and exploiting "hard samples" to boost speaker discrimination, rather than simply increasing overall data volume. Notably, our speaker augmentation approach and conventional data augmentation methods are complementary and can be combined to gain further improvement, though we focus on a deep understanding of speaker augmentation in this paper and so leave the data-and-speaker augmentation as future work.

III. Our Method

A. Speaker Augmentation

1) Step 1: Resampling: We first perform speed perturbation (SP) via time-domain resampling [17]. Given a speech signal $x(t)$, we modify its time axis using a perturbation factor α, resulting in the output signal $y(t)$:

$$y(t) = x\left(\alpha t\right). \tag{1}$$

This time-domain modification induces a corresponding transformation in the frequency domain:

$$X(f) \rightarrow \frac{1}{\alpha} X(\frac{1}{\alpha} f), \tag{2}$$

where $X(f)$ and $\frac{1}{\alpha}X(\frac{1}{\alpha}f)$ denote the Fourier transforms of $x(t)$ and $y(t)$, respectively.

It can be observed that resampling stretches or compresses the spectrogram along both the temporal and the frequency axis. Specifically, the fundamental frequency (F0) and the spectral envelope are raised (downsampling) or lowered (upsampling), leading to changes in pitch and vocal tract characteristics (formants). Consequently, the generated speech retains the same content but exhibits modified speaker traits, simulating a new different speaker.

2) Step 2: Rescaling: To restore the original speech tempo, we apply a time-domain rescaling using the WSOLA algorithm [23] implemented via the 'sox tempo()' function. WSOLA chops the audio into overlapping segments, shifts them in the time domain, and cross-fades them at points where the waveforms are most similar. The entire effect of the WSOLA algorithm is to adjust the speech tempo while maintaining the pitch, thus preserving the speaker traits.

Combining resampling and rescaling, the final augmented speech maintains the content, tempo, and prosody of the original speech while possessing altered speaker traits.

B. Discussion

We argue that this speaker augmentation approach provides two significant advantages for E2E-SE systems.

1) Increased Speaker Diversity: Existing SE benchmark datasets, such as WSJ0-2Mix [18] and LibriMix [19], contain a relatively small number of speakers (typically several hundred), significantly fewer than speaker recognition datasets, which often involve tens of thousands of speakers. With the limited number of speakers in the training data, it is hard to establish the speaker embedding space, which means that speaker generalizability cannot be ensured. By producing pseudo speakers, the speaker embedding space can be better constructed, which improves the speaker encoder's ability to generalize across unseen speakers.

2) Hard Sample Generation: Conventionally, SE models are trained by mixed speech constructed by mixing two randomly selected utterances from two speakers. This naive mixing approach cannot prevent the model from learning spurious cues (e.g., textual similarity) rather than genuine speaker traits. For instance, the model may find that a particular phone sequence can be used to match the enrollment and target speech of the same speaker, and so use that cue to identify the target speech. This risk is particularly high when the number of speakers in the training data is limited, as the speaker generalizability cannot be ensured in this case, so non-speaker cues are likely to be identified as spurious. Unfortunately, SE falls in that case.

Our resampling-and-rescaling approach produces speech that retains the same content, tempo, and prosody as the original utterance and speaker trait is the only changed factor. When mixing the augmented and the original utterances, we

obtain 'hard samples' for which the model can only use speaker traits to extract the target speech. This forces the speaker embedding module and the extractor to capture and utilize genuine speaker characteristics.

In summary, we argue that the effect of speaker augmentation is twofold: more speakers to establish a better generalizable embedding space and increased speaker discrimination by learning from hard samples. We will validate these two augmentations through extensive ablation studies in Section IV.

IV. Experiments

A. Experimental Setup

1) Data: We conduct experiments on two benchmark datasets: WSJ0-2Mix [18] and Libri2Mix [19]. Both datasets are used in their 8kHz versions.

WSJ0-2Mix: This dataset is derived from the Wall Street Journal (WSJ0) corpus [24] and consists of 2-speaker mixtures. The training set comprises 40,000 clean utterances from 101 speakers, while the standard test set includes 3,000 mixtures from 30 speakers.

Libri2Mix: This dataset is based on the LibriSpeech corpus [25] and contains 2-speaker mixtures. We use the *train-100* subset, which consists of 27,800 clean utterances from 251 speakers. Evaluation is performed on two standard test sets: *Libri2Mix clean* (average SNR = 0 dB) and *Libri2Mix noisy* (augmented by noise signals from the WHAM! dataset [26] , average SNR = -2.2 dB), each containing 3,000 mixtures from 40 speakers.

For speaker augmentation, the perturbation factor α was selected from a limited set $\{0.8, 0.9, 1.0, 1.1, 1.2\}$, following the setting in [17]. This expands the number of speakers fivefold. Unless explicitly specified, this fivefold augmentation is the default setting in our experiments.

2) Models: We evaluate our method using two state-of-the-art E2E-SE architectures: DPRNN [20] and SpEx+ [21].

SpEx+: SpEx+ is an E2E-SE model that involves a speech encoder that processes speech signals with three-time scales: $L_1 = 2.5$ ms, $L_2 = 10$ ms, and $L_3 = 20$ ms. The speaker extractor consists of 4 stacked temporal convolutional networks (TCNs) modules, each module containing 8 TCN blocks. The speaker encoder comprises three stacked ResNet blocks, producing 256-dimensional speaker embeddings.

DPRNN: DPRNN is a dual-path RNN-based model initially designed for speech separation. We adopt the DPRNN model from [27] to design our E2E-SE system. which demonstrates strong performance compared to prior E2E-SE system. The speech encoder and decoder follow the configuration presented in [20], with the time scale set to 2.5 ms. The speaker extractor employs a Bi-LSTM network with 128 memory units per direction and a bottleneck size of 64. The speaker encoder consists of three ResNet blocks, generating 256-dimensional speaker embeddings.

The loss function consists of two parts: (1) a speech reconstruction loss based on the scale-invariant signal-to-distortion ratio (SI-SDR) [28], to measure the distortion between the extracted and the clean target speech; (2) a speaker classification loss based on cross-entropy to ensure speaker discrimination.

3) Training setup: We employ a dynamic mixing strategy [29] to generate diverse training samples. For each target utterance x_t, a new mixture y_t is dynamically generated by randomly mixing it with an interfering utterance. In addition, a different utterance e_t from the same target speaker is randomly selected as the enrollment speech. Thus, the training samples are in the form of triplets $\{x_t, e_t, y_t\}$.

For WSJ0-2Mix, the SNR of the mixed speech is uniformly sampled between -5 dB and 5 dB. For Libri2Mix, we follow [19], where the SNR of the 'Libri2Mix clean' set follows a Gaussian distribution with a mean of 0 dB and variance of 16.81 dB, while the SNR of the 'Libri2Mix noisy' set follows a mean of -2.2 dB with a variance of 12.96 dB.

All the models are trained for a maximum of 200 epochs with an initial learning rate of 0.001. The learning rate is reduced by a factor of 0.5 if the validation loss does not reduce for two consecutive epochs. The Adam optimizer is used for training. The source code is publicly available[1].

4) Evaluation metrics: Two evaluation metrics are used to assess model performance.

First, we employ the scale-invariant signal-to-distortion ratio improvement (SI-SDRi) [28] to measure speech extraction quality. A positive SI-SDRi value (SI-SDRi ≥ 0) indicates that the extracted speech is closer to the clean target speech, whereas a negative value suggests that the target speaker's voice is not effectively extracted.

To further evaluate the model's ability to handle target confusion, we adopt the *Negative SI-SDRi Rate* (NSR) [30], defined as:

$$\text{NSR} = \frac{1}{N}\sum_{k=1}^{N}\mathbb{I}(\text{SI-SDRi}^k < 0), \tag{3}$$

where k indexes the test samples, N the total number of test samples, and $\mathbb{I}(\cdot)$ is an indicator function. $\mathbb{I}(\text{SI-SDRi}^k < 0)$ outputs 1 when $\text{SI-SDRi}^k \leq 0$, indicating the occurrence of target confusion in the k-th sample. Otherwise, the indicator function outputs 0.

B. Basic Results

We first evaluate the effectiveness of our proposed speaker augmentation method across the two E2E-SE frameworks and the two benchmark datasets. The experimental results are presented in Table I. It can be observed that our method consistently improves performance under all test conditions across both evaluation metrics, demonstrating its effectiveness.

More interestingly, the effect of speaker augmentation is more pronounced in challenging test conditions. For example, considering the SpEx+ model: in the Libri2Mix clean setting, integrating speaker augmentation relatively improves SI-SDRi by 3.78% and reduces NSR by 6.57%, while in the Libri2Mix

[1]https://github.com/youzhenghai/TSEspkaug

noisy setting, speaker augmentation enhances SI-SDRi by 5.84% and decreases NSR by 21.44%. These results indicate that the generated pseudo-speaker data effectively enhances the model's robustness in complex scenarios. This supports our hypothesis that more speakers in the training data help improve the generalizability of the speaker embedding module.

TABLE I: Performance comparison across different models and datasets with and without speaker augmentation.

Model	Dataset	Setting	SI-SDRi (↑)	NSR (↓)
SpEx+	WSJ0-2Mix	Baseline	16.91	2.35%
		+ SpkAug	17.34	1.10%
	Libri2Mix clean	Baseline	13.23	4.26%
		+ SpkAug	13.73	3.98%
	Libri2Mix noisy	Baseline	10.96	4.85%
		+ SpkAug	11.60	3.81%
DPRNN	WSJ0-2Mix	Baseline	18.62	3.78%
		+ SpkAug	20.03	1.42%
	Libri2Mix clean	Baseline	14.34	4.00%
		+ SpkAug	14.72	3.67%
	Libri2Mix noisy	Baseline	11.45	4.90%
		+ SpkAug	11.98	4.38%

C. Effect of the Number of Augmented Speakers

Next, we investigate how the number of augmented speakers affects the system performance. The results are presented in Table II. One can observe that more pseudo-speakers lead to better performance. In particular, by comparing the results in the 1st row with those in the 5th and 6th rows, we observe that when 125 real speakers are expanded to 250 speakers (half real, half pseudo), the performance is close to that of using 251 real speakers (10.96 vs. 10.82/10.85 in SI-SDRi, 4.85% vs. 4.87%/4.91% in NSR). This further indicates that the pseudo-speakers generated by speaker augmentation can simulate real speakers very well and look sufficient to help establish the speaker embedding space.

TABLE II: Performance of SpEx+ on Libri2Mix noisy with different numbers of (real/pseudo) speakers.

Setting	Spks	α	SI-SDRi (↑)	NSR (↓)
Baseline	251	1.0	10.96	4.85%
+ SpkAug	251 × 3	0.9, 1.0, 1.1	11.42	4.12%
+ SpkAug	251 × 5	0.8, 0.9, 1.0, 1.1, 1.2	11.60	3.81%
Baseline	125	1.0	10.24	6.81%
+ SpkAug	125 × 2	1.0, 1.1	10.82	4.87%
+ SpkAug	125 × 2	0.9, 1.0	10.85	4.91%

D. Effect of the Hard Mixture Samples

We further investigate the effect of hard samples. As discussed in Section III-B, we hypothesize that speaker augmentation generates augmented data that, when mixed with the original utterances, forms more challenging training samples.

To validate this hypothesis, we remove hard samples of different types from the training dataset, making the training set easier:

- **Remove Same Tempo (S.T.) samples**: Augmented speech will only undergo resampling without rescaling, causing a misalignment in tempo.
- **Remove Same Content (S.C.) samples**: Augmented speech will no longer be mixed with the original speech.
- **Remove Same Speaker (S.S.) samples**: Augmented speech will no longer be mixed with speech from the original speaker.

We evaluate the performance changes when different types of hard samples are removed. In our experiments, we fix the total number of training samples and control batch composition to exclude specific hard samples. For S.C., it excludes mixtures of the same content, which account for about 1% and for S.S., it excludes mixtures of the same speaker, which account for about 0.08%. The results are summarized in Table III. Note that we only tested the condition with α set to 0.9 and 1.0, resulting in a twofold speaker expansion.

TABLE III: Effect of hard samples with SpEx+ on Libri2Mix noisy.

Setting	Spks	SI-SDRi (↑)	NSR (↓)
Baseline	125	10.24	6.81%
+ SpkAug	125 × 2	10.85	4.91%
- S.C.	125 × 2	10.82	5.18%
- S.S.	125 × 2	10.65	5.15%
- S.T.	125 × 2	10.81	5.32%
- S.C.	125 × 2	10.77	5.72%
- S.S.	125 × 2	10.57	6.20%

Several key observations can be made:

- Removing any of the three types of hard samples leads to a decline in system performance despite their small proportion, confirming that each contributes positively to model training.
- Among the three, removing S.S. results in the most significant drop in SI-SDRi, while its impact on NSR is less pronounced. Conversely, removing S.T. leads to the highest increase in NSR but has a relatively smaller impact on SI-SDRi. This suggests that different types of hard samples challenge the model in different ways.
- As more hard samples are removed, system performance degrades substantially, supporting our hypothesis that learning from such challenging samples enforces the model to learn genuine and discriminative speaker representations.
- These findings highlight the importance of incorporating diverse hard samples in model training. Further exploration of the underlying effects of each type of hard samples will be considered in our future work.

E. Integration with Metric Learning

Finally, we evaluate the complementarity of our speaker augmentation method with metric learning [10]. Specifically,

in addition to the speech reconstruction loss and speaker classification loss, we add a triplet loss proposed in [10] that aims to minimize the distance between the enrollment embedding and the target speaker embedding, while maximizing the distance between the enrollment embedding and the interfering speaker embedding.

The results are presented in Table IV. It can be seen that combining speaker augmentation and the triplet loss leads to accumulated performance gains. This indicates that speaker augmentation and metric learning are complementary and can be utilized together.

TABLE IV: Performance of SpEx+ by combining speaker augmentation with metric learning.

Dataset	Setting	SI-SDRi (↑)	NSR (↓)
Libri2Mix Clean	SpkAug	13.73	3.98%
	+ Triplet Loss	13.79	3.73%
Libri2Mix Noisy	SpkAug	11.60	3.81%
	+ Triplet Loss	11.67	3.58%

V. Conclusion

In this paper, we propose a simple yet effective speaker augmentation strategy for end-to-end speaker extraction (E2E-SE). By applying a resampling-and-rescaling pipeline, we can generate vast pseudo-speakers, enriching speaker diversity in the training data. Moreover, the generated data preserved the same text, tempo, and prosody of the original utterances, resulting in hard samples that encourage the model to capture and utilize genuine speaker characteristics. We conducted extensive experiments across different model structures and datasets. The experimental results demonstrated that the proposed method consistently enhances SI-SDRi while reducing target confusion. Furthermore, our approach is complementary to metric learning techniques, and their combination leads to additional performance improvement.

We admit the present work needs a more thorough investigation. In particular, the success is based on the present standard benchmark, where the number of speakers is limited, so speaker augmentation is expected to give a substantial contribution. To verify the true value of hard speaker augmentation in learning genuine speaker patterns, we need to test the proposal with larger datasets, e.g., those involving thousands of speakers. This needs to establish a new benchmark for the SE task.

References

[1] 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.

[2] M. Elminshawi, W. Mack, S. R. Chetupalli, S. Chakrabarty, and E. A. Habets, "New insights on target speaker extraction," *arXiv preprint arXiv:2202.00733*, 2022.

[3] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černockỳ, and D. Yu, "Neural target speech extraction: An overview," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, 2023.

[4] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing-Letters and Reviews*, vol. 6, no. 1, pp. 1–57, 2005.

[5] G. R. Naik, W. Wang, *et al.*, "Blind source separation," *Berlin: Springer*, vol. 10, pp. 978–3, 2014.

[6] M. Pal, R. Roy, J. Basu, and M. S. Bepari, "Blind source separation: A review and analysis," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, IEEE, 2013, pp. 1–5.

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

[8] K. Liu, Z. Du, X. Wan, and H. Zhou, "X-SepFormer: End-to-end speaker extraction network with explicit optimization on speaker confusion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] J. Chen, W. Rao, Z. Wang, *et al.*, "MC-SpEx: Towards effective speaker extraction with multi-scale interfusion and conditional speaker modulation," in *INTERSPEECH*, 2023.

[10] Z. Zhao, D. Yang, R. Gu, H. Zhang, and Y. Zou, "Target confusion in end-to-end speaker extraction: Analysis and approaches," in *INTERSPEECH*, 2022, pp. 5333–5337.

[11] Z. Pan, M. Ge, and H. Li, "A hybrid continuity loss to reduce over-suppression for time-domain target speaker extraction," in *INTERSPEECH*, 2022, pp. 1786–1790.

[12] J. Li, K. Zhang, S. Wang, H. Li, M.-W. Mak, and K. A. Lee, "On the effectiveness of enrollment speech augmentation for target speaker extraction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 325–332.

[13] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6109–6113.

[14] C. Deng, S. Ma, Y. Zhang, *et al.*, "Robust speaker extraction network based on iterative refined adaptation," in *INTERSPEECH*, 2021, pp. 3530–3534.

[15] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for

deep speaker embedding.," in *INTERSPEECH*, 2019, pp. 406–410.

[16] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Build a SRE challenge system: Lessons from VoxSRC 2022 and CNSRC 2022," pp. 3202–3206, 2023.

[17] Z. Zhou, S. Xu, S. Yin, L. Li, and D. Wang, "A comprehensive investigation on speaker augmentation for speaker recognition," in *INTERSPEECH*, 2024, pp. 2160–2164.

[18] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 31–35.

[19] 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.

[20] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 46–50.

[21] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li, "Spex+: A complete time domain speaker extraction network," in *INTERSPEECH*, 2020, pp. 1406–1410.

[22] K. Wang, Y. Yang, H. Huang, Y. Hu, and S. Li, "Speakeraugment: Data augmentation for generalizable source separation via speaker parameter manipulation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[23] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 2, 1993, pp. 554–557.

[24] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, *CSR-I (WSJ0) Complete LDC93S6A*, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[26] G. Wichern, J. Antognini, M. Flynn, *et al.*, "Wham!: Extending speech separation to noisy environments," in *INTERSPEECH*, 2019, pp. 1368–1372.

[27] M. Pariente, S. Cornell, J. Cosentino, *et al.*, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *INTERSPEECH*, 2020, pp. 2637–2641.

[28] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 696–700.

[29] A. Alex, L. Wang, P. Gastaldo, and A. Cavallaro, "Data augmentation for speech separation," *Speech Communication*, vol. 152, p. 102 949, 2023.

[30] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and accurate time-domain speaker extraction network," in *INTERSPEECH*, 2020, pp. 1421–1425.