MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Speaker Identity as Sole Supervision for Speech Separation

Boeddeker, Christoph; Masuyama, Yoshiki; Richter, Julius; Edo, Takahiro; Wichern, Gordon; Le Roux, Jonathan

TR2026-137 September 29, 2026

Abstract

Speech separation systems are typically trained with waveformor spectrogram-level reconstruction losses requiring clean references, spatial cues such as multichannel recordings, or remixing strategies. We investigate whether speech separation can instead be trained using speaker identity as the sole supervision signal. We propose a contrastive objective aligning embeddings of separated outputs with embeddings of auxiliary utterances while repelling competing speakers using in-batch negatives. Unlike target speaker extraction, embeddings are used only to define the training objective and are not required at inference, resulting in a conventional speaker-independent separator. Experiments show that speaker-identity supervision alone can train separation systems from scratch to satisfactory performance and that fine-tuning supervised models on noisy mixtures with this objective further improves separation quality.

Interspeech 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Speaker Identity as Sole Supervision for Speech Separation

Christoph Boeddeker, *Yoshiki Masuyama*, *Julius Richter*,
Takahiro Edo, *Gordon Wichern*, *Jonathan Le Roux*

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

{boeddeker,masuyama,richter,tedo,wichern,leroux}@merl.com

Abstract

Speech separation systems are typically trained with waveform- or spectrogram-level reconstruction losses requiring clean references, spatial cues such as multichannel recordings, or remixing strategies. We investigate whether speech separation can instead be trained using speaker identity as the sole supervision signal. We propose a contrastive objective aligning embeddings of separated outputs with embeddings of auxiliary utterances while repelling competing speakers using in-batch negatives. Unlike target speaker extraction, embeddings are used only to define the training objective and are not required at inference, resulting in a conventional speaker-independent separator. Experiments show that speaker-identity supervision alone can train separation systems from scratch to satisfactory performance and that fine-tuning supervised models on noisy mixtures with this objective further improves separation quality.

Index Terms: speech separation, speaker embeddings, contrastive learning, weak supervision

1. Introduction

Speech separation aims to decompose an observed mixture signal into its source signals. Supervised deep learning approaches such as Permutation Invariant Training (PIT) [1, 2], Deep Clustering (DC) [3], and Deep Attractor Network (DAN) [4] have demonstrated strong performance with clean parallel references. However, collecting parallel clean sources in real-world scenarios is not possible and motivated research on training strategies without parallel data.

Unsupervised approaches were first explored in multichannel scenarios, where spatial information can be exploited to enable unsupervised learning. Teacher–student frameworks employing an unsupervised, spatial, neural-free teacher (e.g., spatial mixture models) [5–9], and more recently approaches such as Reverberation As Supervision (RAS), UNSSOR, and Enhanced Reverberation As Supervision (ERAS) [10–13], leverage spatial cues to generate training targets for neural separation models. An obvious key limitation of these methods is that they require multichannel recordings and are therefore not applicable to purely monaural data.

In single-channel setting, Mixture-Invariant Training (MixIT) [14] introduced Mixture of Mixtures (MoM) training, enabling fully unsupervised learning by mixing mixtures, separating them into individual sources, and optimizing a mixture-consistency objective. However, MoM-based approaches are known to suffer from over-separation, which can limit performance and training stability [15, 16]. An alternative approach is self-remixing [15], where mixtures are first separated, remixed within a mini-batch, and separated again before computing a mixture-consistency loss. While the initial proposal [15] relied on initialization from MoM training, follow-up work [17] demonstrated that training from scratch is possible with a carefully designed training schedule.

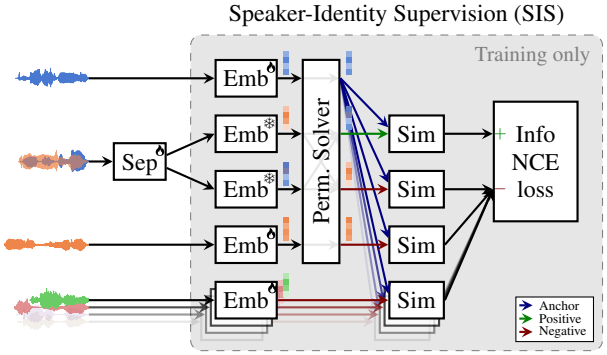

Figure 1: *Overview of the proposed Speaker-Identity Supervision (SIS). "Sep" denotes the separation network and "Emb" the speaker embedding extractor. The embedding extractor is updated only in blocks marked with 🔥 and frozen in blocks marked with ❄. For clarity, only one speaker is shown. The total loss sums over all speakers and batch entries, see Eq.* (7).

In this work, we propose to train a speech separation system using a contrastive objective on speaker embeddings. Instead of optimizing a signal-level reconstruction loss (e.g., time/sample or time–frequency losses), we maximize similarity between embeddings of separated estimates and the corresponding embeddings from auxiliary utterances, while encouraging dissimilarity to embeddings of competing sources. The objective follows a contrastive formulation with in-batch negatives, which are known to improve contrastive representation learning [18, 19]. Our approach, illustrated in Fig. 1, introduces a speaker-identity-based objective into separation training without requiring clean source references or spatial information. Unlike MoM and self-remixing objectives, which operate purely at the signal level, the proposed method trains the separator implicitly via speaker-discriminative embedding comparisons.

Another alternative to signal-level objectives is to use Automatic Speech Recognition (ASR) to supervise separation systems through transcription-based losses [20–22]. These approaches leverage linguistic information, encouraging separated signals to produce correct transcriptions. While this introduces high-level constraints similar in spirit to our work, it requires access to transcriptions and typically relies on pretrained ASR models and/or multichannel recordings. Notably, ASR-based supervision is complementary to the speaker-identity-based ob-

jective considered in this work and could be combined to incorporate both linguistic and speaker-level constraints. Related ideas have also been explored beyond speech separation, for example using sound event detection labels [23] or music transcription labels [24] as weak supervision signals.

A task closely related to speech separation also relying on speaker information is Target Speaker Extraction (TSE) [25, 26], which aims to extract a specific speaker from a mixture given an auxiliary utterance. However, these methods rely on supervised training with clean references and use the auxiliary utterance as a conditioning signal during both training and inference, thus requiring auxiliary data at deployment time. While TSE systems focus on extracting a single target speaker, other approaches aim to extract multiple speakers simultaneously [27, 28], inspired by [29], which proposed a related idea for diarization. In contrast to TSE, we use speaker embeddings only to define the training objective rather than to condition the separator. The resulting model is a conventional speaker-independent separator and does not require auxiliary embeddings at inference time.

Zmolikova et al. [30] explored a related idea in the context of TSE. They adapted a pretrained TSE system directly on the evaluation recordings by exploiting single-speaker segments from long recordings and optimizing a PLDA-based speaker-identity objective combined with a mixture-consistency constraint. This adaptation improves Word Error Rate (WER) and does not require parallel clean references. Their method relies on a pretrained speaker embedding extractor and applies identity-based supervision to refine an already supervised target-conditioned extraction model. In contrast, we investigate whether speaker identity can serve as the primary supervision signal for training a speaker-independent separator, including training from scratch. Unlike TSE, which extracts one speaker at a time, speaker separation jointly estimates all sources. In extraction systems, each output is optimized independently for a known target speaker and is unaware of the estimates of the other speakers. As a result, mixture components such as noise cannot be explicitly assigned to a particular speaker, which implicitly regularizes training. Furthermore, the system of [30] uses speaker identity twice: as an input that conditions the network and as part of the training objective. In contrast, our approach uses speaker identity only in the loss, while the separator itself remains unconditional.

Contributions:

- We investigate training speech separation using speaker identity as the sole supervision signal.
- We introduce a contrastive objective based on speaker embeddings that guides separation without waveform-level reconstruction losses or spatial information.
- Experiments on Libri2Mix show that this supervision can train separators from scratch and can adapt pretrained models to noisy conditions.

2. Speaker Identity as a Supervision Signal

2.1. Setup

Our goal is to train a neural separator $\mathrm{Sep}\,(\cdot)$ that can separate each speech signal from a noisy mixture

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{x}_k + \boldsymbol{n}, \tag{1}$$

of K speakers, where $\boldsymbol{x}_k$ denotes the speech signal of speaker k and $\boldsymbol{n}$ additive noise, leading to estimates

$$\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_K = \mathrm{Sep}\,(\boldsymbol{y}). \tag{2}$$

In addition to the mixture $\boldsymbol{y}$, we assume access at training time, for each speaker present in the mixture, to an auxiliary utterance $\tilde{\boldsymbol{x}}_k$, different from $\boldsymbol{x}_k$, from the same speaker. Such utterances can easily be obtained in realistic conversations, where overlap is sparse and thus leaves regions where only one speaker is active (potentially with noise).

We introduce a neural speaker embedding extractor $\mathrm{Emb}\,(\cdot)$, used at training time only, which maps an utterance to a fixed-dimensional embedding:

$$\hat{\boldsymbol{e}}_k = \mathrm{Emb}\,(\hat{\boldsymbol{x}}_k), \qquad \tilde{\boldsymbol{e}}_k = \mathrm{Emb}\,(\tilde{\boldsymbol{x}}_k). \tag{3}$$

The main idea of our proposed method is to train the separator by jointly optimizing it with the neural speaker embedding extractor such that embeddings of the separated outputs should be similar to those of the auxiliary utterance for the same speaker, while being dissimilar to those of competing sources.

To achieve this goal, we propose to use the contrastive loss function illustrated in Fig. 1 and described below.

2.2. Contrastive loss for Speaker-Identity Supervision (SIS)

To compare speaker embeddings, we use cosine similarity with temperature scaling $\tau = 10$ and flooring at zero:

$$\mathrm{Sim}\,(\boldsymbol{e}_1, \boldsymbol{e}_2) = \max\left(0, \frac{1}{\tau}\frac{\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{e}_2}{\|\boldsymbol{e}_1\|\|\boldsymbol{e}_2\|}\right). \tag{4}$$

To simplify notation, we assume that the separator outputs have been reordered according to the permutation maximizing the total similarity, as in PIT [2], such that

$$\sum_{k=1}^{K} \mathrm{Sim}\,(\tilde{\boldsymbol{e}}_k,\ \hat{\boldsymbol{e}}_k) \ \geq\ \sum_{k=1}^{K} \mathrm{Sim}\left(\tilde{\boldsymbol{e}}_k,\ \hat{\boldsymbol{e}}_{\pi(k)}\right), \quad \forall \pi \in \mathcal{P}_K, \tag{5}$$

where $\mathcal{P}_K$ denotes the set of all permutations of K elements.

We define an InfoNCE loss [18] for speaker k as

$$l(\tilde{\boldsymbol{e}}_k, \hat{\boldsymbol{e}}_k, \mathcal{E}_k) \tag{6}$$
$$= -\log \frac{\exp\left(\mathrm{Sim}\,(\tilde{\boldsymbol{e}}_k, \hat{\boldsymbol{e}}_k)\right)}{\exp\left(\mathrm{Sim}\,(\tilde{\boldsymbol{e}}_k, \hat{\boldsymbol{e}}_k)\right) + \sum_{\boldsymbol{e}\in\mathcal{E}_k} \exp\left(\mathrm{Sim}\,(\tilde{\boldsymbol{e}}_k, \boldsymbol{e})\right)}.$$

Here, the auxiliary embedding $\tilde{\boldsymbol{e}}_k$ acts as the anchor, the estimate embedding $\hat{\boldsymbol{e}}_k$ as the positive example, and $\mathcal{E}_k$ denotes the set of negatives. The negative set $\mathcal{E}_k$ contains the competing estimates $\hat{\boldsymbol{e}}_j$ and auxiliary embeddings $\tilde{\boldsymbol{e}}_j$ for other speakers $j \neq k$ in the mixture, as well as auxiliary embeddings from other mixtures in the same mini-batch, which are included as additional negatives as contrastive objectives benefit from large numbers of negative examples [19].

The total loss is computed as the sum of all InfoNCE losses within a mini-batch

$$\mathcal{L} = \sum_{b}\sum_{k=1}^{K} l(\tilde{\boldsymbol{e}}_k^{(b)}, \hat{\boldsymbol{e}}_k^{(b)}, \mathcal{E}_k^{(b)}), \tag{7}$$

where b indexes the mini-batch.

The formulation in Eq. (4) to Eq. (7) follows the contrastive learning framework of SimCLR [19], with the only modification being the flooring operation in Eq. (4).

2.3. Speaker Embedding Extractor

While the proposed training objective is largely independent of the separator architecture, the design of the speaker embedding extractor is more critical.

The embedder is used only during training and is not part of the inference pipeline, so it does not need to generalize beyond the training distribution. Its sole purpose is to provide informative gradients that guide the separator toward better speaker separation.

One possible design choice is to use a pretrained speaker embedding model. However, speaker verification systems are typically trained to be robust against noise, reverberation, interfering speakers, and distortions introduced by data augmentation. Such augmentations often include degradations that partially resemble separation artifacts. While robustness is desirable for verification, it can be counterproductive in our setting. If the embedder becomes invariant to residual interference or separation artifacts, these artifacts are not penalized by the gradient. Consequently, the separator may retain or even amplify them.

Alternatively, the embedder can be trained jointly with the separator. However, naive joint optimization introduces two challenges. First, since the embedder processes separator outputs, it may adapt to separation artifacts instead of encouraging their removal. Second, common speaker embedding architectures employ batch normalization, which can lead to unintended information exchange across samples within a mini-batch.[1]

To address these issues, we adopt the following strategy. When processing separator outputs, the embedder parameters are frozen: gradients are propagated to the separator parameters, but no gradients are accumulated for the embedder parameters. In a second forward pass, the embedder processes only auxiliary utterances, with gradients enabled to update its parameters. This prevents the embedder from adapting to separator artifacts and mitigates batch-normalization-dependent effects, while still allowing it to learn discriminative speaker representations.

3. Experiments

3.1. Datasets

We consider two experimental scenarios. In the first scenario, we investigate whether speaker identity alone can serve as a supervision signal for separation. Performance is compared to fully supervised training, where the clean source signals generating the mixture are available. In the second scenario, we study a target-domain adaptation setting where we assume access to a supervised source-domain dataset with clean source references available, and a target-domain dataset without such references. Specifically, we consider adapting a separator trained on mixtures of clean data to noisy mixtures using speaker-identity supervision.

For evaluation, we use the Libri2Mix dataset [31] in the 16 kHz max configuration. For the first scenario, we use the `clean` subset (mixtures of two utterances). For the second scenario, we use the `noisy` subset (two utterances mixed with WHAM! noise [32]).

For training, we generate mixtures from LibriSpeech [33] (`train-960`) and WHAM! [32] (`wham_noise_tr`) following the constraints imposed by the proposed Speaker-Identity Supervision (SIS) objective. To prevent information leakage, we split `train-960` into two disjoint subsets: one used for generating mixture signals and one used to provide auxiliary utterances. Without this separation, utterances appearing in a mixture could also be used as auxiliary references, violating the intended setting where ground-truth sources are unavailable.

[1] Observed in preliminary experiments.

3.2. Neural Networks

Separator: For the separator, we employ a comparatively simple architecture to reduce GPU memory consumption during training. This enables larger batch sizes (24 in our experiments due to the limitation from the memory demands of separation models), which are commonly reported to be beneficial for contrastive learning (e.g., [19]).

The network operates on STFT magnitudes and consists of four BLSTMP layers [34] with 256 units. A residual connection is applied around the first two layers. Before the residual connection, the hidden representation is split across $K = 2$ speakers. The last two layers operate on each speaker independently while sharing parameters.

As final non-linearity, we consider two variants: sigmoid and softmax. For the softmax, normalization is performed over the speaker dimension. Since the output is applied as a mask to the STFT signal, this enforces mixture consistency.

For experiments with classical supervision, we use the LogMAE loss [27] with PIT in the time domain.

Speaker Embedding Extractor: For the embedder, we use ECAPA-TDNN[2] [36] in two configurations. The first configuration uses a reduced model size and is trained jointly with the separator. The second configuration employs a pretrained[3] model, which is kept frozen during training.

Our code will be made available online upon acceptance[4].

3.3. Metrics

We evaluate separation performance using signal-level, perceptual, intelligibility, and downstream recognition metrics.

Signal-level quality is measured using BSS Eval SDR [37] in dB. Perceptual quality is assessed with PESQ [38], and speech intelligibility is measured using STOI [39]. To assess the impact on ASR, we compute WER in % using a pretrained NeMo ASR model[5]. Additionally, we report DNSMOS [40], a non-intrusive perceptual quality metric that correlates with human mean opinion scores. For all metrics except WER, higher values indicate better performance, whereas lower WER corresponds to better recognition accuracy.

3.4. Clean Condition

Table 1 shows results on Libri2Mix max clean. As expected, full waveform-level supervision (Wav, C1) yields the strongest performance across all metrics and serves as an upper bound.

Using Speaker-Identity Supervision (SIS) with a jointly trained embedder (C2) achieves 6.5 dB SDR, demonstrating that separation can be learned without waveform- or time-frequency-level reconstruction losses. Prior to experimentation, it was unclear whether optimizing speaker identity alone would preserve linguistic content or induce meaningful waveform separation. The obtained SDR and WER suggest that speaker identity provides an informative training signal which, together with

[2] Implementation from the SpeechBrain [35] toolkit.

[3] `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`

[4] `https://github.com/merlresearch/sis_sep`

[5] `https://huggingface.co/nvidia/stt_en_conformer_ctc_large`

Table 1: *Results on Libri2Mix max (16 kHz) clean (no additive noise). Row Mix corresponds to the unprocessed mixture (no separation). Wav denotes waveform-level supervision (Log-MAE with PIT). SIS denotes the proposed speaker-identity supervision via InfoNCE. Learned indicates that the embedder is trained jointly from scratch, whereas Pretrained refers to the frozen ECAPA-TDNN model. Gray numbers indicate results obtained with access to the clean source signals.*

ID	Embedder	Loss	SDR	PESQ	STOI	WER	DNS MOS
Mix	-	-	–0.1	1.13	0.75	75.0	**2.25**
C1	-	Wav	11.9	1.54	0.89	11.3	2.68
C2	Learned	SIS	**6.5**	**1.28**	**0.80**	30.2	2.11
C3	Pretrained	SIS	1.6	1.18	0.74	**27.7**	1.88

the mixture-consistency constraint imposed by softmax non-linearity, can drive waveform-level separation.

Using a frozen pretrained embedder (C3) substantially reduces SDR (1.6 dB), suggesting that robustness properties of speaker verification models are not well aligned with the proposed objective. Interestingly, WER improves, which may reflect the robustness of modern ASR to residual interference and distortions, even when separation quality is limited.

Overall, the results show that speaker identity alone provides a viable supervision signal for separation and that embedder design is critical for effective training.

3.5. Noisy Condition

Table 2 reports results on Libri2Mix max noisy. The first four rows correspond to the models in Table 1, evaluated under additive WHAM! noise. All clean-trained models degrade substantially, indicating limited robustness under clean-to-noisy domain shift.

Training SIS directly on noisy mixtures (N1, N2) does not consistently outperform clean-only SIS training (C2), and performance depends on the masking non-linearity. With softmax masking (N1), the two-speaker constraint forces noise energy to be assigned to one of the speakers, which preserves mixture consistency but limits denoising potential. Sigmoid masking (N2) relaxes this constraint and enables more flexible noise suppression, which can improve SDR, but may also introduce distortions that adversely affect WER in some configurations.

Initializing from a clean Wav-supervised model provides additional benefits. Fine-tuning with SIS on noisy data (N3, N4) can improve SDR and perceptual metrics (N4) while maintaining reasonable recognition performance. However, the softmax-based mixture-consistency constraint (N3) forces noise energy to be assigned to one of the speakers, which limits denoising capability. Switching to sigmoid masking during fine-tuning (N4) relaxes this constraint and leads to improvements in the signal-level metrics.

The best adaptation result reaches 5.6 dB SDR when combining clean Wav supervision with SIS training on noisy data (N5). While training from scratch serves as a controlled proof-of-concept, initializing from a supervised model reflects a more realistic adaptation scenario. These results indicate that speaker identity-based supervision can effectively adapt a pretrained separator to noisy conditions without sacrificing previously learned separation capability.

For reference, full Wav supervision on noisy data (N6) still provides the strongest overall performance, serving as an upper bound.

Table 2: *Results on Libri2Mix max (16 kHz) noisy (with WHAM! noise). The embedder is trained jointly unless marked with *, which denotes a frozen pretrained ECAPA-TDNN. Clean Loss and Noisy Loss indicate which data subset (clean or noisy) was used for training and which supervision type (Wav or SIS) was applied. Init refers to initialization from the corresponding clean-trained model. Gray numbers indicate results obtained with access to the clean source signals of the noisy mixtures.*

ID	Init	Non-lin.	Clean Loss	Noisy Loss	SDR	PESQ	STOI	WER	DNS MOS
Mix	-	-	-	-	–2.5	1.07	0.66	80.0	1.48
C1	-	Softm.	Wav	-	2.4	1.17	0.71	**39.6**	1.83
C2	-	Softm.	SIS	-	–0.1	1.10	0.63	58.4	1.53
C3	-	Softm.	SIS*	-	–1.2	1.09	0.64	48.1	1.40
N1	-	Softm.	-	SIS	–0.3	1.07	0.63	69.0	1.55
N2	-	σ	-	SIS	1.6	1.16	0.65	62.9	2.09
N3	C1	Softm.	-	SIS	1.5	1.10	0.66	54.2	1.71
N4	C1	σ	-	SIS	3.9	1.20	0.70	56.1	2.36
N5	C1	σ	Wav	SIS	**5.6**	**1.26**	**0.75**	45.3	**2.46**
N6	C1	σ	-	Wav	7.3	1.31	0.79	35.3	2.58

For system N2, informal listening suggested occasional speech-like artifacts not clearly attributable to the input mixture. While such effects are often discussed in the context of generative models, this result indicates that masking-based separators can also exhibit content distortions under weak supervision, particularly when mixture-consistency constraints are relaxed.

4. Conclusion

We proposed a speech separation training strategy that uses speaker identity as the sole supervision signal. Instead of waveform- or time–frequency-level reconstruction losses, the separator is optimized via a contrastive objective on speaker embeddings, which are used only during training.

Experiments on Libri2Mix show that speaker identity alone can induce meaningful separation, yielding positive SDR and usable ASR performance despite the absence of explicit signal reconstruction losses. While performance remains below full waveform supervision, speaker identity-based training provides a viable learning signal, especially when combined with structural constraints such as mixture consistency or when used for fine-tuning to mitigate domain mismatch.

Future work will investigate the interaction between speaker identity-based objectives and structural constraints, as well as combinations with complementary supervision signals such as ASR losses and spatial cues. We also plan to explore stronger separator backbones to better realize the potential of speaker-identity supervision and improve separation quality under weak supervision.

5. Generative AI Use Disclosure

The authors used generative AI tools to assist with language editing and polishing of the manuscript.

6. References

[1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[2] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[4] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE ICASSP*, 2017, pp. 246–250.

[5] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE ICASSP*, 2019, pp. 695–699.

[6] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," in *Proc. IEEE ICASSP*, 2019.

[7] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," in *Proc. IEEE ICASSP*, 2019.

[8] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with Kullback-Leibler divergence based probabilistic loss function," in *Proc. IEEE ICASSP*, 2020, pp. 56–60.

[9] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multichannel source separation," in *Proc. ISCA Interspeech*, 2022.

[10] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, M. Fontaine, and K. Yoshii, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1670–1674, 2021.

[11] R. Aralikatti, C. Boeddeker, G. Wichern, A. Subramanian, and J. Le Roux, "Reverberation as supervision for speech separation," in *Proc. IEEE ICASSP*, 2023.

[12] Z.-Q. Wang and S. Watanabe, "UNSSOR: Unsupervised neural speech separation by leveraging over-determined training mixtures," *Proc. NeurIPS*, vol. 36, pp. 34 021–34 042, 2023.

[13] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. Le Roux, "Enhanced reverberation as supervision for unsupervised speech separation," in *Proc. ISCA Interspeech*, 2024, pp. 607–611.

[14] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Proc. NeurIPS*, vol. 33, pp. 3846–3857, 2020.

[15] K. Saijo and T. Ogawa, "Self-remixing: Unsupervised speech separation via separation and remixing," in *Proc. IEEE ICASSP*, 2023.

[16] J. Kalda, C. Pagés, R. Marxer, T. Alumäe, and H. Bredin, "PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2024, pp. 115–122.

[17] K. Saijo and T. Ogawa, "Remixing-based unsupervised source separation from scratch," in *Proc. ISCA Interspeech*, 2023, pp. 1678–1682.

[18] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020, pp. 1597–1607.

[20] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. IEEE ICASSP*, 2018, pp. 4819–4823.

[21] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *Proc. IEEE ICASSP*, 2020, pp. 7004–7008.

[22] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. IEEE ASRU*, 2019, pp. 237–244.

[23] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2386–2399, 2020.

[24] Y.-N. Hung, G. Wichern, and J. Le Roux, "Transcription is all you need: Learning to separate musical mixtures with score as supervision," in *Proc. IEEE ICASSP*, 2021, pp. 46–50.

[25] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *Proc. ISCA Interspeech*, 2017, pp. 2655–2659.

[26] 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. ISCA Interspeech*, 2019, pp. 2728–2732.

[27] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2024.

[28] B. Zeng, H. Suo, Y. Wan, and M. Li, "Simultaneous speech extraction for multiple target speakers under meeting scenarios," *Journal of Shanghai Jiaotong University (Science)*, pp. 1–7, 2024.

[29] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. ISCA Interspeech*, 2020, pp. 274–278.

[30] K. Žmolíková, 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," in *Proc. ISCA Interspeech*, 2021, pp. 1464–1468.

[31] 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.

[32] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. ISCA Interspeech*, 2019, pp. 1368–1372.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.

[34] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[35] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[36] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. ISCA Interspeech*, 2020, pp. 3830–3834.

[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749–752.

[39] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[40] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE ICASSP*, 2021, pp. 6493–6497.