

---

# Flow Matching-Based Speech Source Separation with Best-of-N Biometric Sampling

---

Anastasia Zorkina<sup>1</sup> Alexandr Anikin<sup>1</sup> Nikita Khmelev<sup>1,2</sup> Anastasiya Korenevskaya<sup>1</sup> Sergey Novoselov<sup>1,2</sup>  
Vladimir Volokhov<sup>1,2</sup> Maxim Korenevsky<sup>2</sup> Yuriy Matveev<sup>1</sup>

## Abstract

Single-channel speech separation remains challenging for real-world deployment due to source permutation ambiguity, sampling variability of generative models, and the difficulty of processing long recordings with chunk-wise inference. We address these issues with a conditional flow-matching-based method that produces an ordered two-source output conditioned on the mixture. A frozen speaker encoder defines the source order during training and is reused at inference for biometric best-of- $N$  candidate selection and chunk-level channel alignment. We evaluate separation quality on Libri2Mix benchmark using SI-SDR, PESQ, and ESTOI, and measure downstream impact using cpWER for automatic speech recognition and EER for speaker verification. The results show that the proposed Transformer U-Net variant is competitive with strong baselines in objective separation metrics and achieves the lowest downstream automatic speech recognition and speaker verification error rates in all evaluated settings.

## 1. Introduction

Speech source separation, also known as the “cocktail party problem”, is a fundamental task in audio processing and speech technologies (Li et al., 2025; Araki et al., 2025; Wang & Luo, 2025). It aims to extract individual speech signals from a mixture, benefiting systems such as automatic speech recognition (ASR), speaker recognition (SR), intelligent voice assistants, and others. Despite rapid progress driven by deep learning, speech separation remains challenging due to acoustic complexity, the ill-posed nature of single-channel inversion, varying overlap patterns, permutation ambiguity, and high speaker variability.

---

<sup>1</sup>ITMO University, Speech Processing Group, Russia <sup>2</sup>Speech Technology Center Ltd., R&D department, Russia. Correspondence to: Anastasia Zorkina <zorkina@speechpro.com>.

Modern deterministic speech source separation systems based on Transformer and Conformer architectures achieve SI-SDRi above 24 dB on the WSJ0-2mix benchmark (Zhao et al., 2024; Shin et al., 2024). Generative counterparts, including hybrid schemes (Lutati et al., 2024; Wang et al., 2024) and fully generative diffusion or flow matching models (Scheibler et al., 2023; Dong et al., 2025; Scheibler et al., 2025), offer new modeling paradigms that improve perceptual naturalness at the cost of sampling variance and higher computation.

Deploying these systems in practice faces additional hurdles: non-causality, processing of long recordings, downstream integration, and computational efficiency. In this work, we propose a practical speech separation system based on conditional flow matching, built on generative speech enhancement models from the NVIDIA NeMo Toolkit (Jukić et al., 2024; Ku et al., 2025). We formulate two-speaker separation as a conditional generation of a structured waveform that contains both separated sources. Besides, we introduce a best-of- $N$  biometric sampling procedure that selects the most speaker-disentangled generation among multiple stochastic candidates.

The main contributions of this work are as follows: 1) A flow-matching-based speech separation method adapted from generative speech enhancement. 2) A best-of- $N$  biometric criterion for inference-time candidate selection. 3) Long-form processing via chunk-wise generation combined with speaker-recognition-based channel tracking. 4) Evaluation of downstream ASR and SR performance after use of our speech source separation system.

## 2. Background

**Speech source separation.** Formally, the task is to recover  $K$  signals  $s_1, \dots, s_K$  from a mixture  $m = \sum_k s_k$ . Permutation-invariant training (PIT) (Yu et al., 2017; Kolbæk et al., 2017) resolves source-ordering symmetry at the loss level but leaves output channels interchangeable at inference, complicating block-wise processing of long-form audio. An information-theoretic upper bound of 23.1 dB SI-SDRi has been established for deterministic separators (Lu-

tati et al., 2022); under heavy overlap, they also introduce perceptual artefacts that SI-SDR does not capture but that downstream ASR/SR penalise (Wang & Luo, 2025). Generative separators sidestep both issues by sampling from  $p(s_1, \dots, s_K | m)$  via diffusion or conditional flow matching (Lipman et al., 2022), at the price of run-to-run sampling variance for which no principled inference-time selection criterion has been proposed.

**Speaker recognition.** Speaker recognition is the task of identifying a speaker from an audio recording. Its canonical formulation is speaker verification (SV): determining whether two recordings originate from the same speaker. State-of-the-art systems are based on deep neural networks, either trained from scratch for this task (e.g., ECAPA-TDNN (Desplanques et al., 2020), ReDimNet (Yakovlev et al., 2024)) or adapted from speech foundation models (e.g., Wav2Vec 2.0-TDNN (Novoselov et al., 2022)). Training is performed as a speaker classification problem (7,205 classes on VoxCeleb 1 & 2 (Nagrani et al., 2017)), employing feature normalization and an additive angular margin in the softmax objective (AAM-Softmax) to enhance class separability. At inference time, the classification layer is discarded, and the model extracts a compact speaker embedding; embeddings from different recordings are then compared via cosine similarity.

**Best-of- $N$  sampling.** Best-of- $N$  returns the highest-ranked of  $N$  independent generator samples under an external verifier, trading inference compute for quality without retraining the generator. It underpins LLM alignment (Stiennon et al., 2020), verifier-based mathematical reasoning (Cobbe et al., 2021), code generation (Chen et al., 2021), and inference-time scaling of text-to-image diffusion (Xie et al., 2025; Choi et al., 2025).

### 3. Proposed Method

#### 3.1. Conditional Flow Matching Formulation

Flow matching defines generation through a time-dependent vector field whose flow map transports samples from a simple prior distribution to the data distribution. Let  $\phi_t$  denote the flow map induced by a vector field  $u_t$ ,

$$\frac{d}{dt}\phi_t(x) = u_t(\phi_t(x)), \quad \phi_0(x) = x. \quad (1)$$

In conditional generation, we approximate this vector field with a neural estimator  $v_t(x_t, c; \theta)$ , where  $x_t$  is the current state,  $c$  is the conditioning information, and  $\theta$  are trainable parameters.

To obtain a tractable training target, we follow the optimal-transport conditional path (Lipman et al., 2022). Given  $x_1 \sim p_{\text{data}}(\cdot | c)$ ,  $x_0 \sim \mathcal{N}(0, I)$ ,  $t \sim \mathcal{U}[0, 1]$ , and an

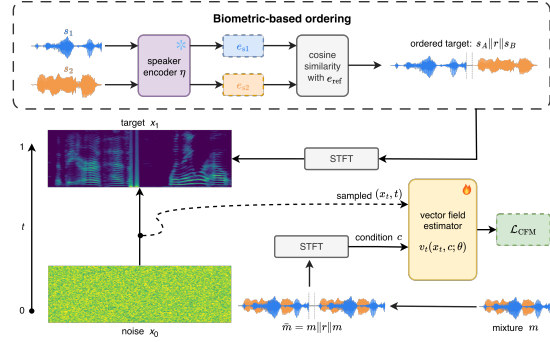


Figure 1. Training procedure of the proposed conditional flow matching demixer.

intermediate state  $x_t$  sampled from this path, the estimator is trained with

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E} \left[ \|v_t(x_t, c; \theta) - u_t(x_t | x_0, x_1)\|_2^2 \right]. \quad (2)$$

We formulate two-speaker speech separation as a conditional generation, guided by some feature  $c$ . The observed mixture waveform is  $m = s_1 + s_2$ , where  $s_1$  and  $s_2$  are speech signals from two speakers. For each mixture,  $c$  is obtained from the complex STFT features of the mixture-derived signal, while  $x_1$  denotes the complex STFT features of the corresponding separated-source signal; therefore, both predicted and target velocities are defined in the same STFT feature space.

#### 3.2. Training Procedure

The training pipeline is shown in Fig. 1. Each training example consists of a mixture  $m$  and two clean sources ( $s_1, s_2$ ); however, these references do not by themselves define a canonical target order. Single-channel separation is ill-posed and permutation ambiguous: the source order and its swapped version define the same mixture, while the model must learn a deterministic output order. To make the target closer to well-defined, we resolve this ambiguity using biometric comparison with a pretrained speaker encoder  $\eta$ : source embeddings are compared with a predefined reference speaker embedding  $e_{\text{ref}}$  using cosine similarity, yielding an ordered pair  $(s_A, s_B)$ , where  $s_A$  is the source closer to  $e_{\text{ref}}$ . We concatenate the ordered sources through a fixed non-speech separator,  $\bar{x} = s_A || r || s_B$ , and duplicate the mixture in the same layout,  $\bar{m} = m || r || m$ ; their complex STFT features define  $x_1$  and  $c$ , respectively. Thus, the source separation task is reduced to a denoising task, where the structural interference of the reversely oriented speech signal  $s_B || r || s_A$  should be removed from  $\bar{m}$ .

During training, we sample  $x_0 \sim \mathcal{N}(0, I)$  and  $t \sim \mathcal{U}(0, 1)$ , then form  $x_t$  using the conditional path above. The neural estimator receives  $x_t$ , the conditioning representation  $c$ , and

the time  $t$ , and is optimized with the conditional flow matching loss. Conditional dropout is applied by replacing  $c$  with zeros with probability  $1 - p_{\text{cond}}$ , exposing the estimator to both conditional and unconditional inputs.

### 3.3. Inference Pipeline

Conventional source separation systems typically operate on a complete input segment, which limits their applicability to long-form recordings. Our inference pipeline instead targets recordings of arbitrary duration by processing the mixture chunk by chunk. This requires two stages: local generation for each chunk and subsequent alignment of chunk-level outputs.

**Chunk-level generation.** Given a mixture recording  $m$ , we split it into 1-second chunks with a 0.5-second hop. Each chunk  $m_c$  is processed independently using the same representation as in training: the mixture-derived condition is encoded into complex STFT features, the initial state is sampled as Gaussian noise  $x_0 \sim \mathcal{N}(0, I)$ , and the learned ODE is solved with an Euler sampler to produce a local source pair  $(\hat{s}_{c,A}, \hat{s}_{c,B})$ .

The stochasticity of flow-based generation allows us to draw multiple plausible separations for the same chunk. We exploit this at inference time through best-of- $N$  biometric sampling: from  $N$  independently generated candidates, we select the pair whose output-channel speaker embeddings are least similar,

$$i^* = \arg \min_i \cos \left( \eta(\hat{s}_{c,A}^{(i)}), \eta(\hat{s}_{c,B}^{(i)}) \right). \quad (3)$$

This favors candidates in which the two generated channels correspond to different speakers.

**Chunk alignment.** After chunk-level generation, the local channel order must be made consistent before reconstruction, since the two generated sources may swap between adjacent chunks. We use biometric alignment based on the speaker embeddings of the generated channels. For each global output channel, embeddings from accepted chunks are accumulated and clustered into two centroids, which gives a more robust channel representation in the presence of noise, leakage, or locally unreliable generations. A new chunk is assigned to the channel order whose embeddings best match the current centroids; otherwise, its two channels are swapped. For controlled analysis, we also report an oracle SI-SDR alignment, which chooses the chunk permutation using ground-truth sources and serves as a reference alignment upper bound. Finally, the aligned chunks are placed back at their original time positions and combined by overlap-add averaging, yielding two full-length separated waveforms.

## 4. Experimental Setup

### 4.1. Data and Preprocessing

We use the two-speaker Libri2Mix benchmark from the LibriMix corpus (Cosentino et al., 2020): 16 kHz max, train-360 for training, and the corresponding test subsets for evaluation. Experiments are reported for the official mix\_clean and mix\_both mixture types.

We additionally apply online augmentation. We load 2-second synchronized mixture/source segments, reverberate both sources with independently selected speaker RIRs, mix them with target-to-interferer SNR sampled from  $[0, 5]$  dB, and add WHAM! (Wichern et al., 2019), MUSAN (Snyder et al., 2015), or CHiME-8 (Cornell et al., 2024) noise to the mixture with SNR sampled from  $[10, 20]$  dB. Targets are the reverberated clean sources. We keep the central 1-second region after augmentation to suppress convolution boundary artifacts.

Targets are represented as  $s_A \| r \| s_B$  and conditions as  $m \| r \| m$ , where  $r$  is a 1000-sample constant separator with amplitude 0.5 (62.5 ms at 16 kHz). Following the NVIDIA NeMo Toolkit implementation (Kuchaiev et al., 2019), we use complex STFT features with hop length 128, magnitude power 0.5, and scale factor 0.33. The FFT length is 254 for NCSN++ (ConvUnet) and 510 for Transformer U-Net (TUnet).

### 4.2. Models and Baselines

We evaluate two velocity estimators: an NCSN++ spectrogram network (Song et al., 2021), referred to as ConvUnet in the results, and a 24-layer Spectrogram Transformer U-Net (TUnet) initialized from the NVIDIA generative speech restoration checkpoint (Ku et al., 2025). Source ordering, best-of- $N$  selection, and chunk alignment use a frozen Wav2Vec 2.0-based speaker embedding model (Novoselov et al., 2022).

Training uses Adam optimizer, learning rate  $10^{-4}$  with cosine decay to  $10^{-6}$ , gradient clipping, and EMA decay 0.999. NCSN++ is trained for up to 300 epochs on NVIDIA A100 GPUs; Transformer U-Net is trained for up to 150 epochs on one A100. We use conditional dropout with  $p_{\text{cond}} = 0.99$ .

Baselines are DiffSep (Scheibler et al., 2023), and SepReformer (Shin et al., 2024) (Large model evaluated in both full-utterance and chunked modes). Additionally recent target speaker extraction method MeanFlow-TSE (Shimizu et al., 2025) was evaluated and compared. For downstream evaluation, we measure automatic speech recognition performance with Whisper V3 (Radford et al., 2023) and speaker recognition performance with Wav2Vec 2.0 (Khmelev et al., 2025) system.

Table 1. Performance comparison of speech source separation and target speaker extraction methods on the Libri2Mix test set in terms of SI-SDR (dB), PESQ, and ESTOI. Values are reported as “clean” / “both”.

Model	SI-SDR	PESQ	ESTOI
ConvUnet (our)	14.42 / 9.28	2.64 / 1.65	0.90 / 0.75
TUnet (our)	17.30 / 11.25	3.11 / 1.92	<b>0.93</b> / 0.79
DiffSep	9.60 / -	2.58 / -	0.78 / -
SepReformer (chunk)	11.30 / -	2.45 / -	0.88 / -
SepReformer (full)	<b>19.22</b> / <b>13.70</b>	3.02 / 2.14	0.92 / <b>0.83</b>
MeanFlow-TSE (full)	17.56 / 11.68	<b>3.27</b> / <b>2.18</b>	0.91 / 0.80

### 4.3. Evaluation Protocols and Metrics

Separation quality is measured in terms of SI-SDR, PESQ, and ESTOI. Downstream ASR quality is measured in terms of cpWER and SV quality is measured in terms of EER. For EER evaluation, we generate verification protocols with clean source utterances as enrollment samples and separated outputs as test samples. For the proposed method all metrics are computed after best-of- $N$  selection with  $N = 4$ .

## 5. Results

Table 1 reports objective separation metrics. In the “clean” setting, TUnet achieves the best ESTOI (0.93) and competitive SI-SDR (17.30 dB), substantially outperforming DiffSep (9.60 dB SI-SDR). SepReformer (full) obtains the highest SI-SDR (19.22 dB), but this configuration processes the full utterance at once, which makes it unsuitable for long recordings in real-world scenarios. In the chunked setting, SepReformer drops to 11.30 dB SI-SDR and 0.88 ESTOI. MeanFlow-TSE achieves the best PESQ (3.27) and the best SI-SDR/PESQ on the “both” subset, but it addresses target speaker extraction and uses target-speaker reference information, making it a different problem setting from blind two-speaker separation. Overall, the proposed ConvUnet and TUnet consistently outperform DiffSep and remain competitive with state-of-the-art approaches across both conditions.

Figure 2 analyzes best-of- $N$  sampling for the proposed TUnet model. Additionally it compares biometric channel selection pipeline with SI-SDR-based oracle selection using ground-truth sources. Increasing  $N$  generally reduces both cpWER and EER for both strategies, with saturation starting around  $N = 4$ . The biometric-based criterion, which does not require ground-truth references, closely approaches the oracle performance, demonstrating its practical utility for inference-time candidate selection in generative separation.

Table 2 reports downstream ASR and SV performance after speech separation on the Libri2Mix clean test set. TUnet achieves the best downstream results among all evaluated systems, with 3.84% cpWER and 0.39% EER. ConvUnet also remains competitive, improving over MeanFlow-TSE and SepReformer (chunk) on both metrics. Al-

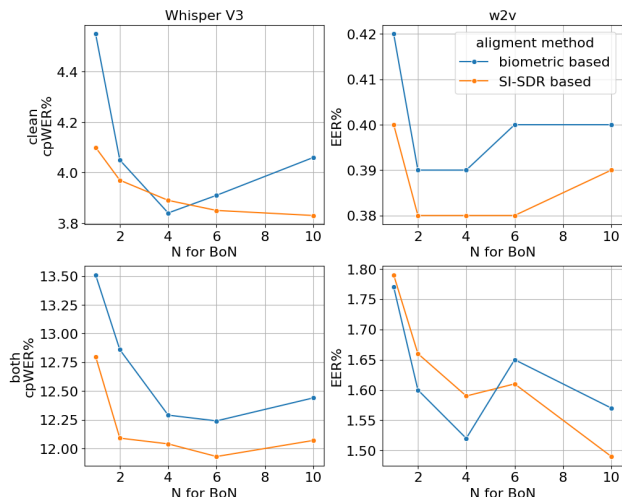


Figure 2. Best-of- $N$  selection on the Libri2Mix test set. Top: “clean”; bottom: “both”. Left: cpWER (%) with Whisper V3 ASR; right: EER (%) with Wav2Vec 2.0 speaker verification. Blue: proposed biometric selection; red: SI-SDR oracle using ground-truth references.

Table 2. Downstream ASR and SV performance on Libri2Mix clean after speech separation. Metrics: cpWER (%) and EER (%).

Model	Whisper V3, cpWER	Wav2Vec 2.0, EER
ConvUnet (our)	4.99	0.51
TUnet (our)	<b>3.84</b>	<b>0.39</b>
MeanFlow-TSE (full)	9.05	2.94
SepReformer (full)	5.02	0.72
SepReformer (chunk)	12.90	1.45

though SepReformer (full) performs close to ConvUnet in the full-utterance setting, its chunk-wise variant degrades substantially, which is consistent with the objective results in Table 1. An additional ablation with alternative downstream backends shows the same trend for Parakeet ASR (Sekoyan et al., 2025) and for ResNet-34 and DistillWhisper-based (Khmelev et al., 2026) SV models, confirming that the observed gains are not specific to a single ASR or SV system. Overall, the downstream evaluation shows that the proposed TUnet is the strongest option in practical chunk-wise processing among the evaluated systems.

## 6. Conclusion

We presented a novel conditional flow matching source separation approach that combines biometric source ordering, best-of- $N$  candidate selection, and chunk-wise processing for long recordings. On Libri2Mix, the Transformer U-Net variant remains competitive in objective separation metrics and achieves the best downstream cpWER and EER among the evaluated systems. These results indicate that the proposed system is well suited for practical speech processing pipelines with downstream ASR and SV tasks.

## References

- Araki, S. et al. 30+ years of source separation research: Achievements and future challenges. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.
- Chen, M. et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Choi, C., Kim, S., and Kim, H. J. Performance plateaus in inference-time scaling for text-to-image diffusion without external models. In *ICML 2025 Workshop on Methods and Opportunities at Small Scale*, 2025.
- Cobbe, K. et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Cornell, S. et al. The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization. In *8th International Workshop on Speech Processing in Everyday Environments (CHiME 2024)*, pp. 1–6, 2024.
- Cosentino, J. et al. LibriMix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*, 2020.
- Desplanques, B., Thienpondt, J., and Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *INTERSPEECH 2020*, pp. 3830–3834, 2020.
- Dong, J., Wang, X., and Mao, Q. EDsep: An effective diffusion-based method for speech source separation. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.
- Jukić, A. et al. Schrödinger bridge for generative speech enhancement. In *INTERSPEECH 2024*, pp. 1175–1179, 2024.
- Khmelev, N. et al. In search of optimal pretraining strategy for robust speaker recognition. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.
- Khmelev, N. et al. Dwhisper-ASV: Distilling knowledge into whisper encoder for fast and domain-robust speaker recognition. In *2026 28th International Conference on Digital Signal Processing and its Applications (DSPA)*, pp. 1–6, 2026.
- Kolbæk, M. et al. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017.
- Ku, P. J. et al. Generative speech foundation model pretraining for high-quality speech extraction and restoration. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.
- Kuchaiev, O. et al. NeMo: A toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.
- Li, K. et al. Advances in speech separation: Techniques, challenges, and future trends. *arXiv preprint arXiv:2508.10830*, 2025.
- Lipman, Y. et al. Flow matching for generative modeling. In *arXiv preprint arXiv:2210.02747*, 2022.
- Lutati, S., Nachmani, E., and Wolf, L. SepIt: Approaching a single channel speech separation bound. *arXiv preprint arXiv:2205.11801*, 2022.
- Lutati, S., Nachmani, E., and Wolf, L. Separate and diffuse: Using a pretrained diffusion model for better source separation. In *International Conference on Learning Representations (ICLR)*, pp. 1–14, 2024.
- Nagrani, A., Chung, J. S., and Zisserman, A. VoxCeleb: A large-scale speaker identification dataset. In *INTER-SPEECH 2017*, pp. 2616–2620, 2017.
- Novoselov, S. et al. Robust speaker recognition with transformers using wav2vec 2.0. *arXiv preprint arXiv:2203.15095*, 2022.
- Radford, A. et al. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.
- Scheibler, R. et al. Diffusion-based generative speech source separation. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.
- Scheibler, R. et al. Source separation by flow matching. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2025.
- Sekoyan, M. et al. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*, 2025.
- Shimizu, R., Jiang, X., and Mesgarani, N. MeanFlow-TSE: One-step generative target speaker extraction with mean flow. *arXiv preprint arXiv:2512.18572*, 2025.
- Shin, U. et al. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52215–52240, 2024.

- Snyder, D., Chen, G., and Povey, D. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.
- Song, Y. et al. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, pp. 1–36, 2021.
- Stiennon, N. et al. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.
- Wang, H. et al. Noise-robust speech separation with fast generative correction. In *INTERSPEECH 2024*, pp. 2165–2169, 2024.
- Wang, Z. and Luo, Z. Speech separation using advanced deep neural network methods: A recent survey. *Big Data and Cognitive Computing*, 9(11):289, 2025.
- Wichern, G. et al. WHAM!: Extending speech separation to noisy environments. In *INTERSPEECH 2019*, pp. 1368–1372, 2019.
- Xie, E. et al. SANA 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025.
- Yakovlev, I. et al. Reshape dimensions network for speaker recognition. In *INTERSPEECH 2024*, pp. 3235–3239, 2024.
- Yu, D. et al. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP 2017 – 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017.
- Zhao, S. et al. MossFormer2: Combining transformer and RNN-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10356–10360, 2024.