НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

КРАТКИЕ СООБЩЕНИЯ

BRIEF PAPERS

doi: 10.17586/2226-1494-2026-26-5-1073-1076
УДК 004.934

Метод одноканального разделения речевых источников для задач распознавания речи и верификации диктора

Анастасия Александровна Зорькина[1]✉, Никита Анатольевич Хмелев[2]

[1,2] ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация
[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] zorkina@speechpro.com✉, https://orcid.org/0000-0002-9434-8897
[2] khmelev@speechpro.com, https://orcid.org/0009-0006-8662-6750

Аннотация
Предложен метод одноканального разделения двух речевых источников на основе условного генеративного моделирования. Рассматриваемая задача характеризуется неопределенностью порядка двух восстановленных речевых сигналов: одна и та же пара разделенных источников может быть представлена в двух перестановках. Для устранения этой неоднозначности используется замороженный кодировщик диктора: он задает порядок источников при обучении, выбирает лучший вариант среди *N* стохастических генераций и согласует источники при обработке записей произвольной длины по фрагментам. На наборе данных Libri2Mix предложенный вариант на основе архитектуры Transformer U-Net конкурентен по фундаментальным метрикам оценки качества и разборчивости обработанной речи и дает наилучшие результаты при использовании в задачах распознавания речи и верификации диктора.

Ключевые слова
разделение речевых источников, генеративное моделирование, сопоставление потоков, распознавание речи, верификация диктора

Ссылка для цитирования: Зорькина А.А., Хмелев Н.А. Метод одноканального разделения речевых источников для задач распознавания речи и верификации диктора // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1073–1076. doi: 10.17586/2226-1494-2026-26-5-1073-1076

Single-channel speech source separation for speech recognition and speaker verification

Anastasia A. Zorkina[1]✉, Nikita A. Khmelev[2]

[1,2] STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation
[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] zorkina@speechpro.com✉, https://orcid.org/0000-0002-9434-8897
[2] khmelev@speechpro.com, https://orcid.org/0009-0006-8662-6750

Abstract
A method for single-channel separation of two speech sources based on conditional generative modeling is proposed. The task is characterized by uncertainty in the order of the two reconstructed speech signals: the same pair of separated sources can be represented in two permutations. To resolve this ambiguity, a frozen speaker encoder is used: it defines the source order during training, selects the best variant among *N* stochastic generations, and aligns the sources during chunk-wise processing of recordings of arbitrary length. On the Libri2Mix dataset, the proposed Transformer U-Net-based variant is competitive in terms of fundamental metrics for evaluating the quality and intelligibility of processed speech and achieves the best results when used in speech recognition and speaker verification tasks.

Keywords
speech source separation, generative modeling, flow matching, speech recognition, speaker verification

© Зорькина А.А., Хмелев Н.А., 2026

For citation: Zorkina A.A., Khmelev N.A. Single-channel speech source separation for speech recognition and speaker verification. *Scientific and Technical Journal of Information Technologies, Mechanics and Optic*s, 2026, vol. 26, no. 5, pp. 1073–1076. doi: 10.17586/2226-1494-2026-26-5-1073-1076

Одноканальное разделение речевых источников является важной задачей для систем автоматического распознавания речи, верификации диктора и речевой аналитики. На практике входной речевой сигнал часто содержит наложение голосов нескольких людей (источников), а ошибки их разделения приводят к уменьшению точности как распознавания речи, так и верификации диктора, поскольку в обработанном речевом сигнале сохраняется остаточная речь собеседника. Исходя из этого, качество разделения речевых источников следует оценивать не только по акустическим метрикам, но и по тому, насколько хорошо после него работают распознавание речи и верификация диктора [1]. Современные нейросетевые разделители часто обучаются с учетом возможных перестановок источников [2]. Однако такой подход устраняет неоднозначность порядка только при вычислении функции потерь и не гарантирует согласованность разделенных сигналов при пофрагментной обработке: один и тот же диктор может оказаться в разных выходных дорожках на соседних фрагментах.

В настоящей работе предложен новый метод разделения наложенной речи двух дикторов, в котором смесь их голосов рассматривается как условие для генерации двух восстановленных раздельных речевых сигналов (рисунок). Задача формулируется в рамках условного моделирования потоков [3]: модель получает спектральные признаки одноканальной записи m, содержащей наложение голосов двух дикторов, и обучается описывать преобразование случайно сгенерированного шума x_0 в целевой разделенный сигнал. Целевой сигнал задается как пара чистых речевых сигналов двух дикторов s_1 и s_2, записанных в одном общем представлении $s_A\|r\|s_B$, где s_A и s_B — речевые сигналы первого и второго дикторов после упорядочивания, а r — фиксированный неречевой разделитель. При этом порядок следования источников определяется с помощью заранее обученного дикторского кодировщика — нейросетевой модели, преобразующей речевой фрагмент в векторное представление диктора. Для каждой смеси векторные представления e_{s_1} и e_{s_2} двух чистых источников сравниваются с заранее заданным фиксированным опорным представлением e_{ref} по косинусной мере близости: более близкий источник записывается первым. Тем самым модель обучается не только удалять взаимную речевую помеху, но и сохранять фиксированную последовательность выходных источников.

На этапе обучения из целевого сигнала формируются его спектральные признаки x_1, а спектральные признаки дублированной смеси $\overline{m} = m\|r\|m$ используются как условие c. Для выбранного момента времени t модель получает промежуточное состояние x_t и условие c и обучается аппроксимировать векторное поле, задающее направление изменения текущего состояния к целевым спектральным признакам разделенных речевых сигналов. Обучение выполняется путем минимизации функции потерь условного моделирования потоков (Conditional Flow Matching, CFM) $\mathcal{L}_{\text{CFM}}$ [3], измеряю-

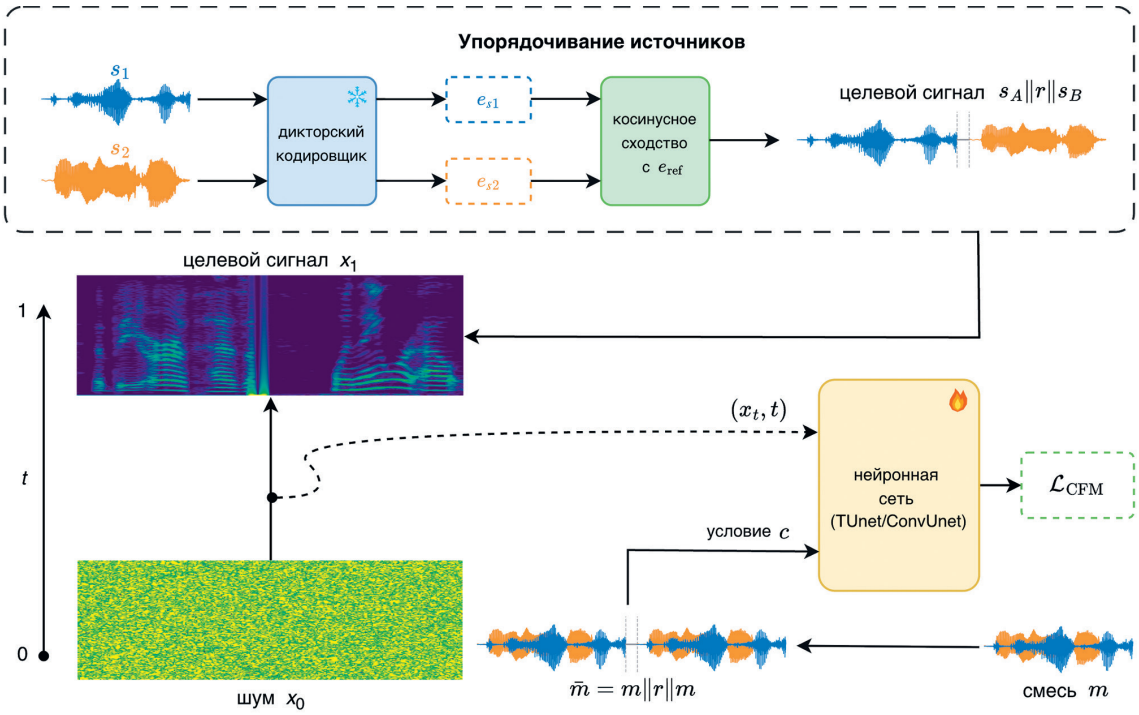


Рисунок. Схема обучения предложенного метода
Figure. Training scheme of the proposed method

щей отклонение предсказанного векторного поля от целевого. В качестве архитектур рассматривались две предобученные модели, ранее применявшиеся в генеративной обработке речи [4]: сверточная сеть NCSN++ (ConvUnet) и Spectrogram Transformer U-Net (TUnet).

На этапе применения запись обрабатывается фрагментами длительностью 1 с с шагом 0,5 с. Для каждого фрагмента генерируется N независимых вариантов разделения, поскольку генеративная модель может выдавать несколько правдоподобных решений для одной и той же смеси. Среди полученных вариантов выбирается тот, в котором косинусное сходство дикторских представлений двух выходных каналов минимально. Такой критерий позволяет выбрать пару сигналов, наиболее вероятно соответствующую двум разным голосам, и не требует доступа к эталонным чистым источникам. После выбора локального результата остается согласовать порядок каналов между соседними фрагментами. В результате для каждой из двух восстановленных аудиодорожек накапливаются дикторские представления уже обработанных фрагментов и вычисляется центроид, описывающий голос соответствующего диктора. Для нового фрагмента оцениваются две возможные перестановки выходных каналов: исходная и обратная. Выбирается та перестановка, при которой дикторские представления каналов имеют наибольшую суммарную косинусную близость к центроидам соответствующих дорожек. После согласования фрагменты возвращаются на исходные временные позиции в соответствующих выходных аудиодорожках, а перекрывающиеся участки усредняются. В итоге формируются две непрерывные аудиодорожки, соответствующие двум разделенным речевым источникам.

Экспериментальная апробация подхода выполнена на наборе данных Libri2Mix [5] для условий с чистыми и зашумленными смесями двух голосов. Сравнение проводилось с моделью SepReformer [6] и методом извлечения целевого диктора MeanFlow-TSE [7]. Качество разделения оценивалось по метрикам Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Perceptual Evaluation of Speech Quality (PESQ) и Extended Short-Time Objective Intelligibility (ESTOI) [1]; влияние на прикладные речевые системы — по метрикам concatenated minimum-permutation Word Error Rate (cpWER) для модели Whisper V3 [8] распознавания речи и Equal Error Rate (EER) для модели dWhisper-ASV верификации диктора [9]. Вариант на основе TUnet при $N = 4$ показал конкурентное качество разделения: на чистых смесях — 17,30 дБ для метрики SI-SDR; 3,11 — PESQ и 0,93 — ESTOI. Влияние разделения дикторов на последующие речевые задачи приведено в таблице. Предложенная схема с TUnet дает наименьшие ошибки среди всех сравниваемых систем: 3,84 % cpWER для распознавания речи и 1,05 % EER для верификации диктора.

Таблица. Влияние разделения речи на ее распознавание и верификацию диктора на чистом наборе данных Libri2Mix

Table. Downstream ASR and speaker verification performance on clean Libri2Mix

Модель	Метрика	
	cpWER, %	EER, %
ConvUnet	4,99	1,28
TUnet	3,84	1,05
MeanFlow-TSE	9,05	3,25
SepReformer	5,02	1,29

Таким образом, при сопоставимом качестве разделения речи предложенный подход оказывается более эффективным в сценариях распознавания речи и верификации дикторов по смеси двух речевых источников (голосов дикторов).

Литература

1. Li K., Chen G., Sang W., Luo Y., Chen Z., Wang S., et al. Advances in speech separation: Techniques, challenges, and future trends // arXiv. 2025. arXiv:2508.10830. doi: 10.48550/arXiv.2508.10830
2. Yu D., Kolbaek M., Tan Z.-H., Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 241–245. doi: 10.1109/icassp.2017.7952154
3. Lipman Y., Chen R.T., Ben-Hamu H., Nickel M., Le M. Flow matching for generative modeling // Proc. of the 11th International Conference on Learning Representations (ICLR). 2023. P. 1–28.
4. Ku P.-J., Liu A.H., Korostik R., Huang S.-F., Fu S.-W., Jukić A. Generative speech foundation model pretraining for high-quality speech extraction and restoration // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025. P. 1–5. doi: 10.1109/icassp49660.2025.10888830
5. Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. LibriMix: An open-source dataset for generalizable speech separation // arXiv. 2020. arXiv:2005.11262. doi: 10.48550/arXiv.2005.11262
6. Shin U.-H., Lee S., Kim T., Park H.-M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation // Proc. of the 38th International Conference on Neural Information Processing Systems. 2024. P. 52215–52240. doi: 10.52202/079017-1655

References

1. Li K., Chen G., Sang W., Luo Y., Chen Z., Wang S., et al. Advances in speech separation: Techniques, challenges, and future trends. *arXiv*, 2025. arXiv:2508.10830. doi: 10.48550/arXiv.2508.10830
2. Yu D., Kolbaek M., Tan Z.-H., Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. doi: 10.1109/icassp.2017.7952154
3. Lipman Y., Chen R.T., Ben-Hamu H., Nickel M., Le M. Flow matching for generative modeling. *Proc. of the 11th International Conference on Learning Representations (ICLR)*, 2023, pp. 1–28.
4. Ku P.-J., Liu A.H., Korostik R., Huang S.-F., Fu S.-W., Jukić A. Generative speech foundation model pretraining for high-quality speech extraction and restoration. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. doi: 10.1109/icassp49660.2025.10888830
5. Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. LibriMix: An open-source dataset for generalizable speech separation. *arXiv*, 2020. arXiv:2005.11262. doi: 10.48550/arXiv.2005.11262
6. Shin U.-H., Lee S., Kim T., Park H.-M. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. *Proc. of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 52215–52240. doi: 10.52202/079017-1655

7. Shimizu R., Jiang X., Mesgarani N. MeanFlow-TSE: One-step generative target speaker extraction with mean flow // arXiv. 2025. arXiv:2512.18572. doi: 10.48550/arXiv.2512.18572
8. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // Proc. of the 40th International Conference on Machine Learning. 2023. P. 28492–28518.
9. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition // Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA). 2026. P. 1–6. doi: 10.1109/dspa69176.2026.11476749

7. Shimizu R., Jiang X., Mesgarani N. MeanFlow-TSE: One-step generative target speaker extraction with mean flow. *arXiv*, 2025. arXiv:2512.18572. doi: 10.48550/arXiv.2512.18572
8. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proc. of the 40th International Conference on Machine Learning*, 2023, pp. 28492–28518.
9. Khmelev N., Anikin A., Korenevskaya A., Zorkina A., Novoselov S., Volokhov V., et al. dWhisper-ASV: Distilling knowledge into Whisper encoder for fast and domain-robust speaker recognition. *Proc. of the 28th International Conference on Digital Signal Processing and its Applications (DSPA)*, 2026, pp. 1–6. doi: 10.1109/dspa69176.2026.11476749

Авторы

Зорькина Анастасия Александровна — младший научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 57807877500, https://orcid.org/0000-0002-9434-8897, zorkina@speechpro.com

Хмелев Никита Анатольевич — научный сотрудник, ООО «ЦРТ-инновации», Санкт-Петербург, 194044, Российская Федерация; аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 58601210700, https://orcid.org/0009-0006-8662-6750, khmelev@speechpro.com

Authors

Anastasia A. Zorkina — Junior Researcher, STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 57807877500, https://orcid.org/0000-0002-9434-8897, zorkina@speechpro.com

Nikita A. Khmelev — Researcher, STC-Innovations LLC, Saint Petersburg, 194044, Russian Federation; PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 58601210700, https://orcid.org/0009-0006-8662-6750, khmelev@speechpro.com

Статья поступила в редакцию 04.07.2026
Одобрена после рецензирования 28.07.2026
Принята к печати 19.09.2026

Received 04.07.2026
Approved after reviewing 28.07.2026
Accepted 19.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»