Проблемы искусственного интеллекта. 2026. N 3 (42). С. 175-186
Problems of Artificial Intelligence. 2026;3(42):175-186.
Системный анализ, управление и обработка информации, статистика
Научная статья

УДК 004.032.26 **DOI** 10.24412/2413-7383-2026-3-42-175-186

В.Г. Шапалин[1], Д. В. Николаенко[2]
Федеральное государственное бюджетное образовательное учреждение высшего образования «Донецкий национальный технический университет»
283001 г. Донецк, ул. Артёма, 58.
[1] shapalinv@gmail.com, ORCID: 0009-0009-9215-9919
[2] dv.nikolaenko@yandex.ru,ORCID: 0009-0006-1806-4952.

ИССЛЕДОВАНИЕ ДАТАСЕТОВ ДЛЯ РАЗДЕЛЕНИЯ ИСТОЧНИКОВ РЕЧИ И ИЗУЧЕНИЕ ВЛИЯНИЯ КРИТЕРИЕВ КАЧЕСТВА ДАННЫХ НА ОБУЧЕНИЕ

V. G. Shapalin[1], D. V. Nikolayenko[2]
Federal stare budgetary educational institution of higher education «Donetsk national technical university». 58 Artema str., Donetsk, 283001.
[1] shapalinv@gmail.com, ORCID: 0009-0009-9215-9919
[2] dv.nikolaenko@yandex.ru,ORCID: 0009-0006-1806-4952.

ANALYSIS OF SPEECH SOURCE SEPARATION DATASETS AND THE IMPACT OF DATA QUALITY CRITERIA ON MODEL TRAINING

Данная статья исследует датасеты для разделения источников речи, оценивая объективные критерии качества и разнообразия данных, а также оценивает влияние каждого критерия на эффективность обучения, что может быть полезным для дальнейшей разработки или модификации данных для работы с речью. Были рассмотрены базовые критерии качества данных, а также характеристики, которые можно оценить статистически. Для оценки качества разделения, датасеты были обучены на нейросетевых моделях, и протестированы на общем тестовом наборе данных. Была проведена оценка влияния медианной составляющей и разброса в каждом критерии на качество разделения, а также с помощью количественной оценки корреляции в виде критерия Пирсона. Итоговый результат показал, что сильнее всего качеством модели положительно коррелирует средний процент речи и перекрытия двух спикеров.
Ключевые слова: датасет, разделение источников звука, качество данных, критерий Пирсона, клиппинг, LUFS.

This paper investigates datasets for speech source separation by evaluating objective criteria of data quality and diversity, and by analyzing the impact of each criterion on training performance, which may be useful for the further development or modification of speech datasets. Both basic data quality criteria and characteristics that can be statistically measured are considered. To evaluate separation quality, the datasets were used to train three neural network models and were tested on a common evaluation dataset. The influence of the median component and the variability of each criterion on separation quality was assessed, together with a quantitative correlation analysis using the Pearson correlation coefficient. The final results showed that the strongest positive correlations with model quality were the average proportion of speech and the overlap percentage between two speakers.
Key words: dataset, sound source separation, data quality, Pearson correlation coefficient, clipping, LUFS.

Введение

Разделение источников речи (*speech separation*) является одной из ключевых задач обработки аудиосигналов, находящейся на стыке цифровой обработки сигналов и глубокого обучения. Современные нейросетевые архитектуры демонстрируют хорошие результаты при восстановлении сигнала, однако качество итогового разделения в значительной степени определяется не только моделью, но и характеристиками исходных данных. Особенность построения данных, отношение громкостей спикеров, их перекрытие и наличие клиппинга влияют на обучаемость модели и обобщающую способность на новых данных.

Настоящая работа посвящена анализу ряда критериев качества данных для задач разделения источников речи и исследованию их влияния на эффективность моделей разделения. Разделение источников речи является актуальной и не до конца решенной проблемой – в сложных акустических условиях разделение проблематично, что делает актуальной и задачу потенциального улучшения качества данных. Научная новизна работы заключается в слабоизученности данной темы: критерии качества данных для обработки речи являются слишком обобщенными, и могут плохо подходить для задачи разделения источников речи, а исследований корреляции критериев аудиоданных и качества обучения задаче разделения источников речи нет совсем. Целью исследования является определение наиболее значимых показателей, их количественная оценка и анализ взаимосвязей этих показателей с метриками качества разделения. По полученным результатам будут сформированы итоговые выводы и рекомендации по дальнейшему улучшению аудиоданных для построения более эффективных систем разделения речи.

Смежные исследования

Существует ряд работ, посвященных качеству речевых аудиоданных и данных для обучения нейросетей в задачах обработки речи, однако специализированных исследований, ориентированных именно на критерии качества датасетов для разделения источников речи, не обнаружено. В [1] проанализированы 69 аудиодатасетов, описаны их ключевые характеристики, а также проведен PCA анализ эмбеддингов. При этом все рассматриваемые наборы данных ориентированы на задачи аудио языкового моделирования (ALM) и преобразования аудио в текст, что делает полученные результаты полезным фоном, но слабо применимыми к задаче разделения источников речи. В [2] изложена методология повышения качества речевых корпусов, основанная на построении чистых и хорошо контролируемых наборов аудиоданных. Однако предлагаемые критерии качества данных носят обобщенный характер и преимущественно ориентированы на задачи распознавания и обработки речи и естественного языка (ASR и NLP), а не на специализированные задачи разделения источников речи. В [3] исследуется влияние масштаба обучающего корпуса и параметров модели на качество распознавания речи. Показано, что при увеличении объема данных качество модели возрастает, причем на относительно малых датасетах этот эффект выражен сильнее. Вместе с тем в [3] рассматривается именно задача ASR, и полученные зависимости напрямую не переносятся на задачу разделения источников речи.

Критерии качества данных

В рассмотренных выше работах анализируются, в основном, общие критерии качества данных, которые носят достаточно обобщенный характер. Так как речевые корпусы технически являются базами данных, на них распространяются те же

требования – «Данные должны быть доступными, точными, полными, согласованными, актуальными, полезными, валидными и т.д.». Данные о создании датасета должны быть задокументированы и доступны для обычного пользователя.

К общим параметрам можно отнести формат аудиоданных – их кодек (или его отсутствие), частоту дискретизации и канальность. Важно не использовать форматы с потерей данных – необходимо выбирать кодеки без потерь или не использовать их вообще. К общим параметрам также можно отнести размер корпуса. Для задач распознавания речи в [4] показано, что прирост объема данных особенно заметен на относительно малых датасетах. Дальнейшее увеличение данных уменьшает прирост качества обучения экспоненциально. Тем не менее, применимость этого наблюдения к задаче разделения источников речи требует отдельной проверки.

Вторая категория критериев данных – это общие особенности речи говорящих – их возраст, пол, акценты, а также языки, на которых записана речь. Также сюда можно отнести манеру речи – спокойная и эмоциональная речь воспринимается по-разному, и одно может содержать фонетические признаки, не характерные для другого. Отдельное внимание уделяется шуму. Для задачи разделения источников речи, шум может присутствовать в смешанном сигнале вместе со спикерами. Умеренное присутствие шума может быть полезно для повышения эффективности разделения в сложных условиях. Нужно учитывать, что шум должен быть различной природы [5].

Взяв наиболее релевантные критерии, а также добавив другие, потенциально значимые, был составлен общий список для объективной оценки качества данных по разделению источников речи. В первую категорию входят обязательные критерии, без удовлетворения которых данные считаются «грязными», что сразу делает такие наборы «невалидными». К этим критериям относятся:

- Одинаковая частота дискретизации;
- Одинаковое количество каналов;
- Одинаковая длина в сэмплах (для каждой пары смесь-источники).

Остальные данные подразумевают их количественную оценку. Предполагается, что чем разнообразнее эти критерии, тем более качественно будет происходить разделение на тестовых данных. К таким критериям относится:

- Наличие клиппинга у смеси и источников;
- Громкость:
  - Разброс уровня громкости у смесей;
  - Соотношение громкости одного источника к другому;
  - Отношение громкости смеси к шуму.
- Паузы и пересечения:
  - Доля пауз между репликами у спикера;
  - Процент перекрытия двух спикеров.

Рассматриваемые датасеты для изучения

Для исследования были выбраны четыре набора данных:

- LibriMix [6];
- LombardGrid2Mix [7], [8];
- EchoSet [9];
- LRS2-2Mix [10].

Для упрощения постановки задачи рассматривались только моноканальные датасеты со смесями из двух спикеров. Отдельно необходимо уточнить о генерации датасета LombardGrid2Mix: ввиду трудностей с генерацией шумных записей в оригинальном репозитории, был разработан модифицированный вариант, который незначительно отличается в плане генерации громкости шума [8].

Методика исследования и его результаты

Для изучения вышеуказанных критериев был разработан набор скриптов на языке Python. Исходный код указан в публичном репозитории на github [11]. Так как объемы данных могут быть очень большими, данные ограничены 160000 аудиофайлов, что является достаточной выборкой для обобщения на всех данных. Данные выбираются псевдослучайно, с заранее заданным сидом для воспроизводимости. Результаты по каждому датасету записываются в txt-файл, а затем выводятся в общий график.

– **Длина, частота дискретизации и количество каналов.** Все 3 критерия можно объединить в одну категорию. Если частота дискретизации и количество каналов хотя бы в одном файле будет отличаться, это не даст возможность для обучения. Это также относится к длине файла – она должна быть одинакова внутри каждой пары смесь-спикеры. При невыполнении этих критериев, на таких данных нельзя обучать. В ходе изучения 4 вышеуказанных наборов данных, было доказано, что они удовлетворяют всем 3 критериям.

– **Процент клиппинга.** Клиппинг – это искажение звука, которое появляется, когда амплитуда сигнала выходит за диапазон, который может передать или записать система. Различают «жесткий» (hard) и «мягкий» (soft) клиппинг. В первом случае, сигнал резко обрезается, оставляя плоскую вершину. Во втором случае, сигнал затухает постепенно, и сигнал получается более плавным, но звучание все еще будет неестественным [12].

Так как мягкий клиппинг преобладает в аналоговом сигнале, было решено его опустить, так как изучался только цифровой сигнал. Результат работы показан в таблице 1.

Таблица 1 – Процент клиппинга в датасетах

Датасет	% клиппинга в источниках	% клиппинга в смесях
LibriMix	0 %	0 %
LombardGRID2mix	0 %	0 %
EchoSet	1.92 %	4.74 %
LRS2-2Mix	0 %	0 %

– **Разброс уровней громкости в источниках, отношение громкостей источников, отношение громкости смеси к шуму.** Здесь были рассмотрены разброс уровней громкости всех источников, отношение громкости одного источника к другому (внутри одной смеси) и отношение громкости смеси к громкости шума. В качестве объективной метрики громкости был использован loudness units relative to full scale или LUFS, использующий стандарт ITU-R BS.1770. Он учитывает особенности восприятия громкости человеком на разных частотах, благодаря чему является наиболее популярным инструментом измерения громкости. Статистическая оценка этих данных описана в таблицах 2, 3 и 4. Min, max – минимум и максимум, P_{25}, P_{75} – перцентили 25% и 75% соотвественно, μ – математическое ожидание, σ – стандартное отклонение. Некоторые из представленных датасетов не имеют исходного источника-шума, поэтому громкость смеси к громкости шума для них вычислить невозможно.

Таблица 2 – Громкость смесей в LUFS: статистика по датасетам

Датасет	Min	P_{25}	Медиана	P_{75}	Max	μ	σ
LibriMix	-31.07	-26.49	-25.31	-24.33	-20.96	-25.44	1.49
LombardGRID 2mix	-24.73	-17.32	-16.24	-15.18	-10.79	-16.3	1.6
EchoSet	-67.89	-27.73	-24.4	-21.44	-6.59	-24.95	5.43
LRS2-2Mix	-25.53	-17.74	-16.64	-15.66	-11.51	-16.78	1.58

Таблица 3 – Отношение LUFS для двух источников: статистика по датасетам

Датасет	Min	P_{25}	Медиана	P_{75}	Max	μ	σ
LibriMix	0	1.09	2.36	4.02	12.36	2.7	1.91
LombardGRID2mix	0	1.26	2.51	3.79	11.43	2.58	1.57
EchoSet	0	1.62	3.26	5.09	38.07	3.82	3.37
LRS2-2Mix	0	2.51	5.03	7.53	18.6	5.07	2.99

Таблица 4 – Отношение LUFS чистой смеси к шуму: статистика по датасетам

Датасет	Min	P_{25}	Медиана	P_{75}	Max	μ	σ
LibriMix	-6.28	5.8	7.93	10.09	31.31	7.92	2.94
LombardGRID2mix	-3.37	5.14	9.75	14.95	33.8	9.93	5.65

– **Паузы и пересечения спикеров.** Для определения наличия речи в аудиофрагменте использовалась нейросеть Silero-VAD для детекции голосовой активности. Модель возвращает метки начала и конца речевых интервалов в сэмплах. На их основе формируется бинарная маска наличия речи (массив размерности, равной числу сэмплов, где 1 соответствует наличию речи, а 0 — ее отсутствию), после чего вычисляется процент речевой активности в каждом аудиофайле [13], [14]. Результат исследования показан в таблицах 5 и 6.

Таблица 5 – Процент речи в файлах-источниках: статистика по датасетам

Датасет	Min	P_{25}	Медиана	P_{75}	Max	μ	σ
LibriMix	13.74	80.91	86	90.27	99.99	84.92	7.64
LombardGRID2mix	0	88.4	91.52	93.33	100	90.68	4.16
EchoSet	0	58.07	67.07	76.73	97.83	67.2	12.95
LRS2-2Mix	18.3	95.1	96.7	98.3	100	95.61	5.14

Таблица 6 – Процент перекрытия двух спикеров: статистика по датасетам

Датасет	Min	P_{25}	Медиана	P_{75}	Max	μ	σ
LibriMix	12.11	68.82	75.24	80.98	98.45	74.38	9.42
LombardGRID2mix	0	85.64	88.03	91.03	99.91	87.82	4.72
EchoSet	0	41.93	51	60	91.97	50.58	13.69
LRS2-2Mix	0	93.5	95.1	96.7	100	93.16	6.58

Сравнение эффективности обучения 3 моделей разделения источников речи на датасетах

Для проверки эффективности обучения, были использованы 3 различных типа моделей для разделения источников речи:

- Сверточная модель Conv-TasNet [15];
- Рекуррентная модель Dual Path RNN [16];
- Трансформерная модель DPTNet [17].

Чтобы данные для обучения меньше зависели от тестовых, был разработан самостоятельный тестовый набор данных, содержащий по 1.6 часов аудио из датасетов Librimix, LombardGrid-2mix и LRS2-2mix, что дает суммарную длину в 4.8

часов. Все данные взяты из тестового набора каждого из датасетов. Выборка из каждого тестового набора псевдослучайна, с заранее заданным сидом. Чтобы ускорить обучение и снизить влияние размера датасета, для обучения и тестирования использовалось ограниченное количество смесей: 7500 аудиозаписей, длина каждой из которых ограничена 2 секундами. Обучающая и тестовая выборки формировались псевдослучайно. Длительность обучения ограничена 40 эпохами.

Результаты обучения моделей на различных датасетах приведены в таблице 7. В качестве функции потерь при обучении всех трех моделей использовалась метрика SI-SDR [18], которая рассчитывается по формуле 1.

$$\mathrm{SI-SDR} = 10\log_{10}\frac{\|\alpha s\|^2}{\|\hat{s}-\alpha s\|^2},\ \alpha = \frac{\langle \hat{s}, s\rangle}{\|s\|^2},$$

где s – эталонный сигнал,

$\hat{s}$ – оцененный сигнал.

Таблица 7 – Сравнение эффективности работы 3 разных типов моделей при обучении на различных датасетах. Метрика качества обучения SI-SDR.

Датасет / Модель	LibriMix	LombardGrid 2Mix	EchoSet	LSR2-2Mix
Conv-TasNet	4.9500	3.3977	1.8261	3.7988
Dual path RNN	3.4054	4.6398	-0.4545	4.6541
DPTNet	5.2441	4.5607	-0.3820	5.8454
Среднее	4.5332	4.1994	0.3298	4.7661

Поиск корреляции между данными и производительностью нейросети.

Диаграммы рассеяния зависимости SI-SDR от медианных значений и стандартных отклонений выбранных показателей по датасетам представлены в соответствии с рисунками 1 и 2. По оси X расположены соответствующие значения, по оси Y – качество разделения по метрике SI-SDR. Значения коэффициента корреляции Пирсона приведены в таблицах 8 и 9. С учетом крайне малого числа наблюдений (четыре датасета) статистические выводы следует интерпретировать с осторожностью, даже при формальном уровне значимости $p<0.05$.

Коэффициент корреляции рассчитывался между четырьмя значениями медиан или стандартных отклонений выбранных показателей по датасетам и усредненным значением SI-SDR по всем трем моделям. Корреляция отношения громкостей смесь-шум невозможно подсчитать из-за ограниченности данных.

Стоит отдельно отметить, что ввиду малого набора данных (4 элемента), и отсутствию детерминированности при обучении, критерий Пирсона нельзя считать достаточно надежным инструментом, но его результаты можно использовать для дополнительной, более объективной интерпретации данных.

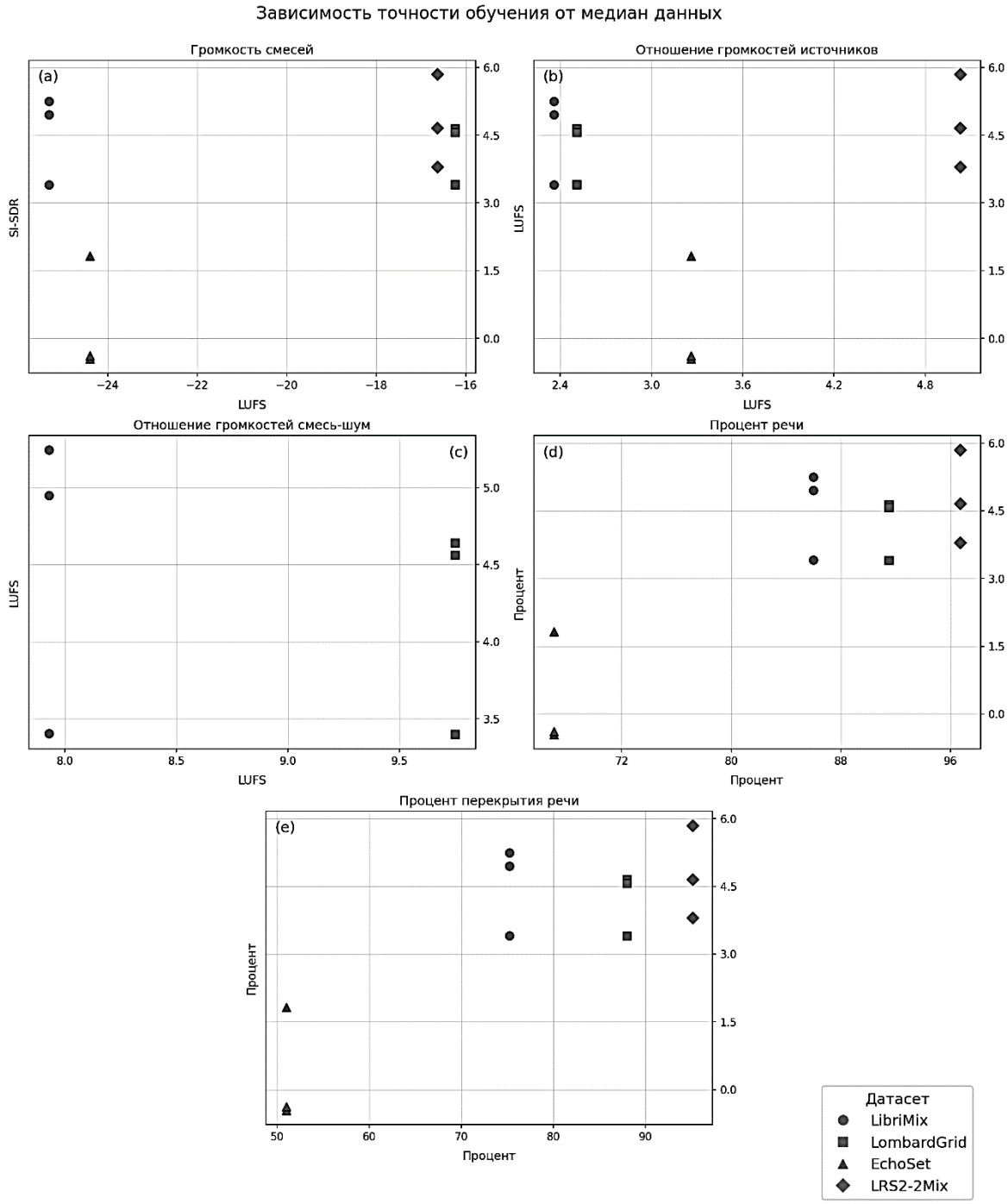


Рисунок 1 – Диаграмма рассеяния зависимости качества разделения (метрика SI-SDR) от медианного значения данных по датасетам. График (a) – распределение громкостей смесей, (b) – отношение громкостей двух источников, (c) – отношение громкости смеси к громкости шума, (d) – процент речи в аудио, (e) – процент перекрытия двух спикеров

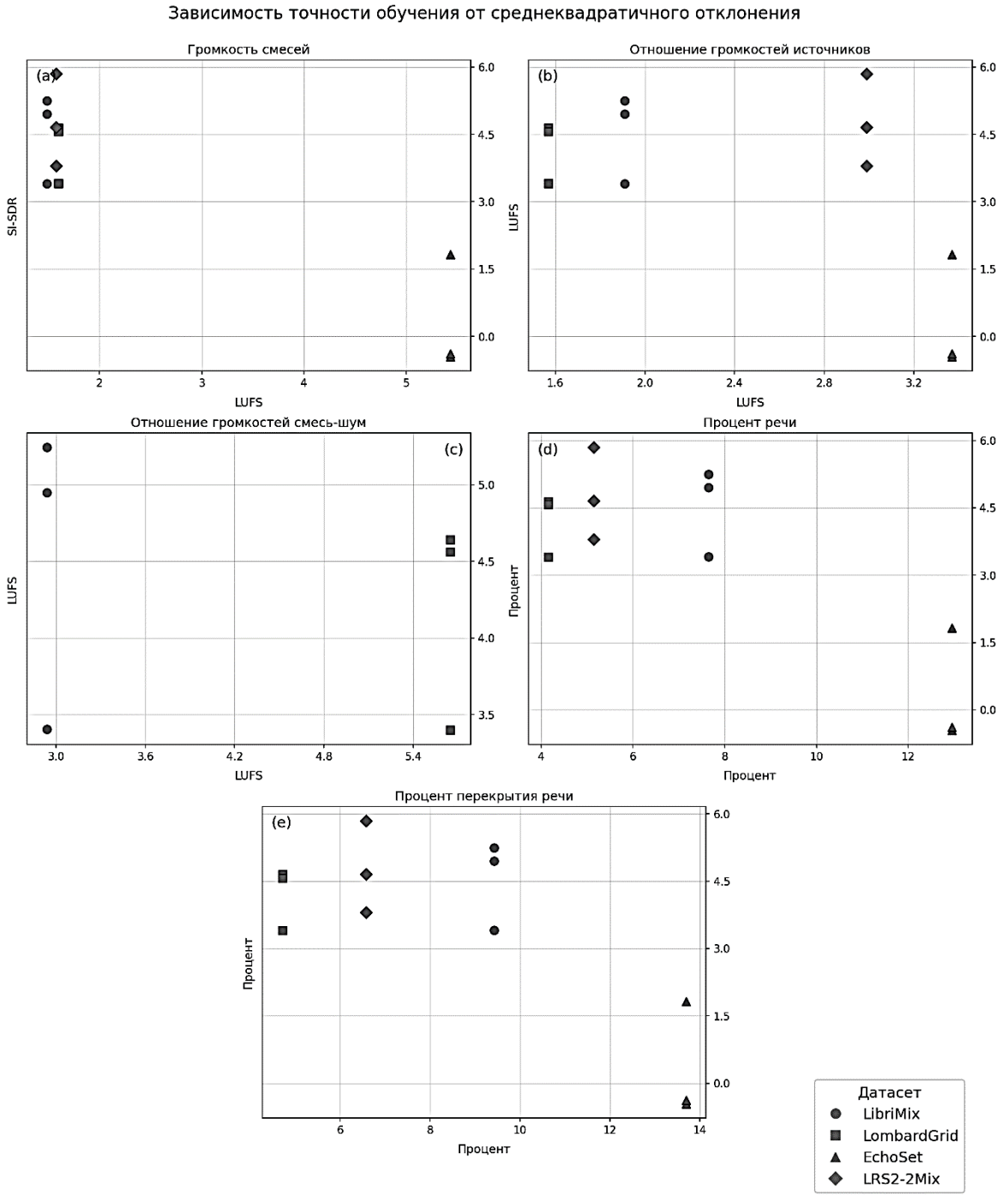


Рисунок 2 – Диаграмма рассеяния зависимости качества разделения (метрика SI-SDR) от среднеквадратичного значения данных по датасетам. График (a) – распределение громкостей смесей, (b) – отношение громкостей двух источников, (c) – отношение громкости смеси к громкости шума, (d) – процент речи в аудио, (e) – процент перекрытия двух спикеров

Таблица 8 – Корреляция между качеством разделения данных и их медианными значениями

Критерий	Критерий Пирсона	
	r	p
Распределение громкостей	0.4966	0.5034
Отношение громкостей источников	0.1027	0.8973
Процент речи в аудио	0.9501	0.0499
Процент перекрытия двух спикеров	0.9121	0.0879

Таблица 9 – Корреляция между качеством разделения данных и их стандартным отклонением

Критерий	Критерий Пирсона	
	r	p
Распределение громкостей	-0.9943	0.0057
Отношение громкостей источников	-0.6312	0.3688
Процент речи в аудио	-0.9071	0.0929
Процент перекрытия двух спикеров	-0.8369	0.1631

Заключение

В ходе работы по изучению датасетов по разделению источников речи было изучено влияние ряда объективных факторов на качество обучения. Следует отметить, что из-за малого количества данных по полученным результатам можно сделать лишь ограниченные и осторожные выводы:

1. Наличие клиппинга в датасете, вероятно, не улучшает качество разделения данных, а, скорее всего, ухудшает.
2. Больший средний процент речи в аудиофайлах-источниках положительно коррелирует с итоговым результатом.
3. Аналогичная тенденция наблюдается и для процента перекрытия двух источников. Однако из-за малого числа наблюдений надежность этих оценок ограничена, и соответствующие выводы следует рассматривать как предварительные. Это может объясняться малым количеством данных. Одним из вариантов уточнения данного вопроса является искусственное увеличение датасетов (например, разделение одного датасета на несколько с контролируемым значением медианного значения процента перекрытия речи).
4. Для разброса распределения громкостей наблюдается отрицательная корреляция: чем меньше разброс, тем выше качество обучения. Вместе с тем столь сильная связь может частично объясняться малым числом датасетов, поэтому ее желательно уточнить, разделив исходные наборы на поднаборы с контролируемыми параметрами, как предложено в п. 3.
5. В разбросе отношения громкостей источников EchoSet явно выбивается из общего тренда, что предположительно объясняется наличием больших хвостов, увеличивающих среднеквадратическое отклонение. Рекомендуется повторить эксперимент, ограничив данные между 10 и 90-ми процентилями.

Наиболее значимым критерием оказалось большое среднее значение процента речи. Менее интуитивным является корреляция между качеством разделения и минимальным разбросом уровней громкости. Данный критерий предлагается исследовать

в дальнейшем, разделив данные на большее количество отдельных наборов данных. В целом можно предположить, что полученные результаты могут быть полезны для дальнейшего совершенствования данных не только в задачах разделения источников речи, но и в смежных задачах обработки речи и звука. Однако, ввиду ограниченности наблюдений, полученные выводы требуют дальнейшей проверки и уточнения. Например, с помощью деления больших датасетов на малые поднаборы данных с контролируемыми статистическими параметрами. Результаты работы могут служить основой для дальнейшего, более детального изучения критериев качества данных в рассматриваемой области.

Список источников

1. Audio-Language Datasets of Scenes and Events: A Survey / Wijngaard G., Formisano E., Esposito M., Dumontier M. // IEEE Access. 2025. Vol. 13. P. 20328–20360.
2. Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora / A. Wieczorkowska // Applied Sciences. 2025. Vol. 15, no. 4. 1848 p.
3. Droppo J. Scaling Laws for Acoustic Models / J. Droppo, O. Elibol // arXiv preprint arXiv:2106.09488. 2021.
4. Robust Speech Recognition via Large-Scale Weak Supervision / Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. // ICML'23: Proceedings of the 40th International Conference on Machine Learning. 2023. Pp. 28492-28518.
5. Сайгин А. Подготовка данных для обучения нейронных сетей, решающих задачу разделения источников сигнала / А. Сайгин, Н.Плотникова // E-Scio. № 3 (78). 2023. С.112-122.
6. LibriMix: An Open-Source Dataset for Generalizable Speech Separation / Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. // arXiv preprint arXiv:2005.11262. 2020.
7. Ewert I., Borsdorf M., Li H., Schultz T. Does the Lombard Effect Matter in Speech Separation? Introducing the Lombard-GRID-2mix Dataset. / I. Ewert, M. Borsdorf, H. Li, T. Schultz // Proc. Interspeech 2024, 2024. Pp. 577-581.
8. LombardGrid2Mix modification [Электронный ресурс]. URL: https://github.com/ShapalinVitaliy/LombardGRID2mix_mod (дата обращения: 13.04.2026).
9. TIGER: Time-frequency Interleaved Gain Extraction and Reconstruction for Efficient Speech Separation / Xu M., Li K., Chen G., Hu X. // arXiv preprint arXiv:2410.01469. 2026.
10. Gao R. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency / R. Gao, K. Grauman // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021. Pp. 15490-15500.
11. SepDataCheck: набор скриптов для анализа датасетов для разделения источников речи [Электронный ресурс]. URL: https://github.com/ShapalinVitaliy/SepDataCheck.
12. A Survey and an Extensive Evaluation of Popular Audio Declipping Methods / Záviška P., Rajmic P., Ozerov A., Rencker L. // IEEE Journal of Selected Topics in Signal Processing 15, 2020. Pp. 5-24.
13. Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier [Электронный ресурс]. GitHub repository, 2024. URL: https://github.com/snakers4/silero-vad (дата обращения: 13.04.2026).
14. Савенков И. Разработка VAD-алгоритма на основе глубокого обучения / И. Савенков, Т. Ермоленко, А. Цыбик // Проблемы искусственного интеллекта. 2022. № 1 (24). С. 65-74.
15. Luo Y. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation / Y. Luo and N. Mesgarani // IEEE/ACM Trans. Audio, Speech and Lang. Proc. 27, 8 (August 2019), 2019. P. 1256–1266.
16. Luo Y. Dual-path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation / Y. Luo, Z. Chen, T. Yoshioka // 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 46–50.
17. Chen J. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation / J. Chen, Q. Mao, D. Liu // ArXiv preprint arXiv:2007.13975. 2020.
18. SDR – Half-baked or Well Done? / Roux J. L., Wisdom S., Erdogan H., Hershey J.R. // ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019. Pp. 626-630.

References

1. Audio-Language Datasets of Scenes and Events: A Survey / Wijngaard G., Formisano E., Esposito M., Dumontier M. // IEEE Access. – 2025. – Vol. 13. – P. 20328–20360.
2. Wieczorkowska A. Methodology for Obtaining High-Quality Speech Corpora / A. Wieczorkowska // Applied Sciences. – 2025. – Vol. 15, no. 4. – 1848 p.
3. Droppo J. Scaling Laws for Acoustic Models / J. Droppo, O. Elibol // arXiv preprint arXiv:2106.09488. 2021.
4. Robust Speech Recognition via Large-Scale Weak Supervision / Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. // ICML'23: Proceedings of the 40th International Conference on Machine Learning. – 2023. – Pp. 28492-28518.

5. Saygin A. Data preparation for training neural networks for the task of signal source separation / A. Saygin, N. lotnikova // E-Scio. – 2023. – № 3 (78). – С. 112-122.
6. LibriMix: An Open-Source Dataset for Generalizable Speech Separation / Cosentino J., Pariente M., Cornell S., Deleforge A., Vincent E. // arXiv preprint arXiv:2005.11262. 2020.
7. Ewert I. Does the Lombard Effect Matter in Speech Separation? Introducing the Lombard-GRID-2mix Dataset / I. Ewert, M. Borsdorf, H. Li, T. Schultz // Proc. Interspeech 2024. – P. 577-581.
8. LombardGrid2Mix modification [Электронный ресурс]. – URL: https://github.com/ShapalinVitaliy/LombardGRID2mix_mod (дата обращения: 13.04.2026).
9. TIGER: Time-frequency Interleaved Gain Extraction and Reconstruction for Efficient Speech Separation / Xu M., Li K., Chen G., Hu X. // arXiv preprint arXiv:2410.01469. 2026.
10. Gao R. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency / R. Gao, K. Grauman // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). – Nashville, TN, USA, 2021. – P. 15490-15500.
11. SepDataCheck: a set of scripts for analyzing datasets for speech source separation [Электронный ресурс]. – URL: https://github.com/ShapalinVitaliy/SepDataCheck.
12. A Survey and an Extensive Evaluation of Popular Audio Declipping Methods / Záviška P., Rajmic P., Ozerov A., Rencker L. // IEEE Journal of Selected Topics in Signal Processing. – 2020. – № 15. – P. 5-24.
13. Silero Team. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. GitHub repository [Электронный ресурс]. – 2024. – URL: https://github.com/snakers4/silero-vad (дата обращения: 13.04.2026).
14. Savenkov I. Development of a VAD algorithm based on deep learning / I. Savenkov, T. Ermolenko, A. Tsybik // Problems of Artificial Intelligence. – 2022. – № 1 (24). – P. 65-74.
15. Luo Y. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation / Y. Luo and N. Mesgarani // IEEE/ACM Trans. Audio, Speech and Lang. Proc. 27. – 8 (August 2019), 2019. – P. 1256–1266.
16. Luo, Y. Dual-path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation / Y. Luo, Z. Chen, T. Yoshioka // 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2020. – P. 46–50.
17. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation / J. Chen, Q. Mao, D. Liu // ArXiv preprint arXiv:2007.13975. – 2020.
18. SDR – Half-baked or Well Done? / Roux J. L., Wisdom S., Erdogan H., Hershey J.R. // ICASSP 2019 - 2019 IEEE International Conference on Acoustics. – Speech and Signal Processing (ICASSP). – 2019. – P. 626-630.

RESUME

V.G. Shapalin, D.V. Nikolaenko

Analysis of speech source separation datasets and the impact of data quality criteria on model training

Despite the active development of speech processing, including tasks such as recognition, extraction, and source separation, there is still too little information on the principles of constructing speech corpora and datasets—particularly regarding what is required to achieve better speech processing quality beyond basic recommendations.

A separate issue is the evaluation of data quality for speech source separation. Although a number of studies have investigated speech data, they remain too generalized to serve as a basis for developing a clear set of criteria defining what data for speech separation should be like.

For this work, only quantitative criteria were selected – those that can be objectively evaluated and expressed through measurable indicators. The entire study was designed around the possibility of objectively assessing data quality criteria, which excludes a number of factors that are either too abstract or impossible to calculate without the necessary metadata, such as age, gender, accent, and similar attributes.

The final results indicate that a higher proportion of speech within individual audio recordings and a greater overlap between two sources exhibit the strongest correlation with data quality. However, conclusions regarding the diversity of noise levels, source loudness levels, and the overall mixture cannot be drawn with the same level of confidence, despite the presence of partial relationships. To obtain more accurate results, it may be necessary to develop algorithms for dividing the entire dataset into controlled subsets with fixed variance and median parameter values.

РЕЗЮМЕ

В.Г. Шапалин, Д.В. Николаенко
Исследование датасетов для разделения источников речи и изучение влияния критериев качества данных на обучение

Несмотря на активное развитие направления обработки речи, в том числе в таких задачах, как распознавание, выделение и разделение источников, слишком мало данных об особенностях построения речевых корпусов данных – о том, что необходимо для лучшего качества обработки речи, не считая базовых рекомендаций.

Отдельно стоит вопрос оценки качества данных для разделения источников речи. Несмотря на ряд работ, исследующих речевые данные, они слишком обобщенные, чтобы на их основе можно было разработать ряд критериев, какими должны быть данные для разделения речи.

Для данной работы были выбраны именно количественные критерии, которые могут быть объективно оценены, и иметь количественные показатели. Все исследование строилось на том, чтобы можно было объективно оценить критерии качества данных, что отсекает ряд критериев, которые или слишком абстрактные, или которые невозможно подсчитать без необходимых метаданных: возраст, пол, акцент и т.д.

Итоговые результаты показывают, что большее наличие речи в отдельно аудио и большее пересечение двух источников больше всего коррелирует с качеством данных. Однако, нельзя с такой же уверенностью делать выводы по разнообразию громкостей шума, источников и смеси в целом, несмотря на частичную взаимосвязь. Для более точных результатов, возможно, потребуются разработать алгоритмы для разделения целого набора данных на контролируемые поднаборы с фиксированными параметрами дисперсии и медианных значений.

Шапалин Виталий Геннадиевич – аспирант, ДонНТУ, 283001 г. Донецк, ул. Артёма, 58. Телефон +7(949) 995-20-24, shapalinv@gmail.com. *Область научных интересов*: нейронные сети, цифровая обработка звука, обработка речи. ORCID: 0009-0009-9215-9919
Николаенко Денис Владимирович – Кандидат технических наук, доцент, доцент кафедры "Компьютерная инженерия", декан факультета «Интеллектуальные системы и программирование», ДонНТУ, 283001 г. Донецк, ул. Артёма, 58. Телефон +7(949) 356-13-90, dv.nikolaenko@yandex.ru. *Область научных интересов*: Интеллектуальные и робототехнические системы в транспорте. ORCID: 0009-0006-1806-4952.

Shapalin Vitaliy Gennadiyevich – Postgraduate Student, Federal stare budgetary educational institution of higher education «Donetsk national technical university». 58 Artema str., Donetsk, 283001, phone +7(949) 995-20-24, shapalinv@gmail.com, Research interests: neural networks, digital sound processing, speech processing. ORCID: 0009-0009-9215-9919
Nikolayenko Denis Vladimirovich – PhD in engineering sciences, associate professor, associate professor of computer engineering department, dean of faculty “Intelligent systems and programming”, Federal stare budgetary educational institution of higher education «Donetsk national technical university». 58 Artema str., Donetsk, 283001, phone +7(949) 356-13-90, dv.nikolaenko@yandex.ru. Research interests: intelligent and robotic systems in transport. ORCID: 0009-0006-1806-4952.

Статья поступила в редакцию 17.06.2026;
одобрена после рецензирования 08.07.2026; принята к публикации 02.09.2026
The article was submitted 17.06.2026; approved after reviewing 08.07.2026;
accepted for publication 02.09.2026.