2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

doi 10.26089/NumMet.v27r213

УДК 517.968;
519.642.3

Исследование алгоритмов онлайн-диаризации спикеров для задачи автоматического распознавания речи

А. В. Полевой
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
Москва, Российская Федерация
ORCID: 0009-0000-2216-0468, e-mail: polevoianton@bk.ru

Н. В. Лукашевич
Московский государственный университет имени М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
Научно-исследовательский вычислительный центр,
Москва, Российская Федерация
ORCID: 0000-0002-1883-4121, e-mail: louk_nat@mail.ru

Аннотация: Несмотря на успехи в области автоматического распознавания речи, существует большое количество сценариев (спонтанная речь, смесь дикторов, перебивания и др.), в которых современные системы работают нестабильно, не позволяя качественно передать смысл сказанного. Задача диаризации спикеров, состоящая в разбиении аудиопотока на сегменты, относящиеся к разным говорящим, остается одной из наиболее сложных и актуальных проблем в обработке речи в режиме реального времени. Особенно трудными являются случаи длительных многоспикерных записей, которые типичны для форумов, корпоративных мероприятий и панельных дискуссий. Существующие потоковые решения часто ограничены по числу спикеров, требуют высоких вычислительных ресурсов или обладают значительной задержкой при обработке звукового потока. В данной работе представлен алгоритм каскадной обработки потокового распознавания речи с онлайн-диаризацией, состоящий из детектора голосовой активности (VAD), модели извлечения эмбеддингов и онлайн-кластеризации на основе Probabilistic Spherical Discriminant Analysis (PSDA). Для повышения стабильности в потоковом режиме предложены эвристики устойчивости, уменьшающие количество ложных переключений спикеров и обеспечивающие согласованное поведение модели при ограниченном временно́м контексте. Результаты, полученные при помощи предложенного каскадного алгоритма онлайн-диаризации спикеров, значительно превосходят результаты существующих систем на собственном русскоязычном наборе многоспикерных данных. Также демонстрируется эффективность каскадных подходов для потоковой диаризации в условиях ограниченных вычислительных ресурсов.

Ключевые слова: онлайн-диаризация, русская речь, PSDA, потоковая обработка, эмбеддинги речи, PCA, VAD.

Для цитирования: Полевой А.В., Лукашевич Н.В. Исследование алгоритмов онлайн-диаризации спикеров для задачи автоматического распознавания речи // Вычислительные методы и программирование. 2026. **27**, № 2. 194–207. doi 10.26089/NumMet.v27r213.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

Research of online speaker diarization algorithms for the task of automatic speech recognition

Anton V. Polevoi
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Moscow, Russia
ORCID: 0009-0000-2216-0468, e-mail: polevoianton@bk.ru

Natalya V. Loukachevitch
Lomonosov Moscow State University,
Faculty of Computational Mathematics and Cybernetics,
Research Computing Center,
Moscow, Russia
ORCID: 0000-0002-1883-4121, e-mail: louk_nat@mail.ru

Abstract: Despite the advances in the field of automatic speech recognition, there are a large number of scenarios (spontaneous speech, an overlapping of speakers, interruptions, etc.) in which modern systems work unstable, failing to accurately convey the intended meaning of the utterance. The task of speaker diarization, which involves splitting an audio stream into segments corresponding to individual speakers, remains one of the most difficult and relevant problems in real-time speech processing. Particularly challenging are cases of prolonged multi-speaker recordings, which are typical for forums, corporate events, and panel discussions. Existing streaming solutions are often limited in the number of speakers, require high computing resources, or have a significant delay in processing the audio stream. This paper presents an algorithm for cascaded processing of streaming speech recognition with online diarization, consisting of a voice activity detector (VAD), an embedding extraction model, and online clustering based on Probabilistic Spherical Discriminant Analysis (PSDA). To increase reliability in streaming mode, stability heuristics are proposed that reduce the number of false speaker switches and ensure consistent model behavior under a limited temporal context. The results obtained using the proposed cascade algorithm for online speaker diarization significantly outperform those of existing systems based on own Russian-language multi-speaker dataset. The effectiveness of cascade approaches for stream diarization in conditions of limited computing resources is also demonstrated.

Keywords: online speaker diarization, Russian speech, PSDA, streaming processing, speech embeddings, PCA, VAD.

For citation: A. V. Polevoi, N. V. Loukachevitch, "Research of online speaker diarization algorithms for the task of automatic speech recognition," **27** (2), 194–207 (2026). doi 10.26089/NumMet.v27r213.

---

1. Введение. Процесс диаризации спикеров, представляющий собой разделение аудиозаписи по говорящим путем определения того, "кто и когда говорил", является базовым элементом систем распознавания речи, анализа совещаний и мультимедийного поиска. При этом перекрытия голосов, шум, изменчивость акустических условий и большая продолжительность записи затрудняют выделение спикеров в реальном времени. Общая постановка задачи заключается в простановке временны́х меток начала и окончания речи для каждого диктора.

Особую сложность представляет задача *онлайн-диаризации* спикеров, где алгоритм должен обрабатывать аудиопоток в реальном времени без доступа к будущим данным. Это критично для приложений потоковой обработки, в особенности для систем синхронного перевода в онлайн-режиме. Основные вызовы онлайн-диаризации включают: 1) необходимость принятия решений при ограниченном временно́м контексте; 2) минимизацию задержки обработки; 3) обеспечение стабильности при длительных потоках с неограниченным числом спикеров; 4) эффективное использование вычислительных ресурсов.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

Современные подходы к диаризации делятся на два основных класса:

1. *Каскадные системы.* Включают последовательную обработку через детектирование речи (VAD), извлечение эмбеддингов (нормализованных векторных представлений) и этап кластеризации. Такие системы являются модульными, интерпретируемыми и могут работать на CPU, но требуют тщательной настройки компонентов.
2. *Сквозные (end-to-end) архитектуры.* Для них характерна единая нейросеть, обучаемая на всей задаче целиком. Решения, такие как LS-EEND [1], Sortformer [2], демонстрируют высокую точность, однако требуют мощных GPU, ограничены длиной сегмента (до 10–15 мин) и числом спикеров (обычно не более 4), что затрудняет применение подобных архитектур для длительных многоспикерных записей.

Основной вклад данной работы заключается в следующем:

1. Предложен надежный каскадный алгоритм онлайн-диаризации при работе с неограниченным числом спикеров. Он реализован в рамках единого подхода к автоматическому распознаванию речи с использованием модели Streaming FastConformer [3].
2. Для обеспечения устойчивости разработаны эвристики, уменьшающие количество ложных переключений спикеров и гарантирующие согласованное поведение модели при ограниченном временно́м контексте потоковой обработки.
3. Создан русскоязычный тестовый набор данных RusAudioForum для публичных мероприятий с высококачественной разметкой, включающий записи известных форумов и конференций.
4. Проведена комплексная оценка метода на стандартных наборах данных (AMI [4], VoxConverse [5]) и собственном датасете.

Настоящая работа демонстрирует, что каскадные подходы остаются конкурентоспособными при решении задач потоковой диаризации длительных многоспикерных записей, обеспечивая баланс между точностью, задержкой и требованиями к вычислительным ресурсам.

2. Предпосылки и обзор подходов.

2.1. Классические методы и каскадные подходы. Ранние системы диаризации опирались на GMM–HMM модели и агломеративную кластеризацию на основе так называемых i-векторов [6]. Под i-векторами (identity vectors) [7] понимают компактное векторное представление речевого сигнала, отражающее характеристики спикера. Работа с векторами для решения задачи диаризации основана на факторном анализе и статистическом моделировании с использованием универсальной фоновой модели (Universal Background Model, UBM) и гауссовых смесей (GMM). Обычно реализуемая как смесь гауссовых распределений статистическая модель UBM обучается на большом наборе речевых данных от различных говорящих и служит общим эталоном для сравнения характеристик спикеров. Эффективное извлечение признаков спикера из акустических статистик речевого сегмента возможно благодаря тому, что i-векторы моделируют вариативность между разными спикерами через низкоразмерное факторное подпространство. Каждый речевой сегмент представляется фиксированным по размеру вектором (размерности 400), который кодирует как характеристики спикера, так и различные вариации (канал записи, шум и др.). Применение i-векторов широко распространено в задачах верификации и идентификации спикеров [7], а также диаризации [8], где они используются в сочетании с вероятностными методами классификации, такими как Probabilistic Linear Discriminant Analysis (PLDA).

С развитием глубокого обучения появились подходы на основе d-векторов [9] — векторных представлений спикеров, извлекаемых с помощью глубоких нейронных сетей (DNN), обученных на задаче идентификации спикеров. В отличие от i-векторов, базирующихся на статистическом моделировании, d-векторы извлекаются как активации последнего скрытого слоя DNN, обученной классификации спикеров [9]. Метод использует подход обучения с учителем, где сеть обучается отличать различных спикеров, после чего последний слой становится универсальным эмбеддингом спикера. Процедура извлечения d-векторов обычно включает усреднение активаций по временно́й оси речевого сегмента, что обеспечивает инвариантность к длительности сегмента. Развитие представлений о d-векторах привело к созданию более эффективных архитектур, таких как x-векторы, использующих TDNN (Time-Delay Neural Networks) и современные подходы с вниманием, такие как ECAPA–TDNN [10]. Современные методы обучения, подобные Generalized End-to-End (GE2E) loss [11], позволяют обучать модели для извлечения более качественных эмбеддингов, лучше различающих спикеров.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

Помимо классического i-векторного подхода с PLDA, в последние годы активно развиваются методы, адаптированные для работы с нормированными эмбеддингами (например, на основе архитектур типа ECAPA–TDNN), которые лежат на единичной гиперсфере $\mathbb{S}^{d-1}$.

Сферический PLDA (constrained PLDA) представляет собой адаптацию стандартного PLDA для нормированных эмбеддингов, где гауссовские распределения заменяются на распределения фон Мизеса–Фишера (von Mises–Fisher, VMF), заданные на единичной гиперсфере. В модели предполагается, что эмбеддинги спикера и отдельные его реализации генерируются VMF-распределениями с общим центром $\boldsymbol{\mu}$ и двумя параметрами концентрации: межспикерским b и внутриспикерским w. Данный подход был подробно исследован в работе [12], где предложена эффективная процедура обучения, адаптированная для задач онлайн-распознавания и кластеризации спикеров.

В то же время в ряде работ эмбеддинги спикеров для диаризации разделяют по косинусному расстоянию и пороговым значениям, без явного вероятностного моделирования (далее по тексту cosine similarity embedding aggregation, CSEA). Для каждого спикера хранится список его эмбеддингов. Вычисленный эмбеддинг для текущего спикера сравнивается со средним эмбеддингом, накопленным на предыдущих шагах. Если максимальная схожесть превышает или равна пороговому значению (которое подбирается эмпирически), эмбеддинг приписывается соответствующему спикеру. Если схожесть меньше порогового значения, создается новый спикер.

Современные пакеты программного обеспечения, такие как `pyannote.audio` [13], реализуют модульный каскадный подход с использованием глубоких нейронных сетей для извлечения эмбеддингов. Однако в большинстве случаев они предназначены для офлайн-анализа с доступом ко всему аудиофайлу.

Каскадные системы обычно состоят из трех этапов: 1) детектирование голосовой активности для выделения речевых сегментов; 2) извлечение эмбеддингов спикеров с помощью нейронных сетей (ResNet, ECAPA–TDNN [10]); 3) кластеризация эмбеддингов для группировки сегментов по спикерам. Преимущества такого подхода включают модульность, интерпретируемость и возможность работы на CPU, что делает его привлекательным для практических приложений.

2.2. Онлайн-диаризация и потоковые методы. В онлайн-режиме алгоритм не имеет доступа к данным на всем аудиофайле целиком, поэтому прогнозы модели должны быть устойчивыми. Малейшие вариации тона или шума в исходном сигнале не должны вызывать ложные смены спикеров.

Типичные потоковые подходы, такие как UIS-RNN [14] и Diart [15], обеспечивают хорошее качество, но требуют крупных временны́х промежутков (2–5 с) для накопления достаточного контекста, что увеличивает задержку обработки. Модуль Diart демонстрирует ухудшение качества при малых буферах и требует использования GPU.

2.3. Ограничения существующих решений. Несмотря на большое количество подходов, существующие решения имеют ряд ограничений:

1. *Ограничение по числу спикеров.* Большинство сквозных методов (FS-EEND, Sortformer) рассчитаны не более чем на 4 спикера, что недостаточно для панельных дискуссий и форумов.
2. *Ограничение по длительности.* Сквозные методы работают с сегментами до 10–15 мин, что затрудняет обработку длительных записей (30–60 мин и более).
3. *Высокие требования к ресурсам.* Сквозные методы требуют мощных GPU, что осложняет их применение в ресурсоограниченных сценариях.
4. *Задержка обработки.* Многие онлайн-методы требуют накопления значительного контекста (2–5 с), что увеличивает задержку и может быть неприемлемо для приложений реального времени.
5. *Недостаточная стабильность.* При длительных потоках наблюдается деградация качества из-за накопления ошибки.

Эти ограничения затрудняют практическое использование существующих методов для реальных мероприятий длительностью 60 минут и более с переменным числом спикеров.

3. Метрики. Для количественной оценки качества решения задачи диаризации спикеров используется стандартная метрика Diarization Error Rate (DER). Данная метрика отражает долю времени, в течение которого система диаризации допускает ошибки при определении активности и идентичности говорящих.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

Формально DER определяется следующим образом:

$$\mathrm{DER} = \frac{T_{\mathrm{FA}} + T_{\mathrm{MISS}} + T_{\mathrm{ERR}}}{T_{\mathrm{TOTAL}}},$$

где T_{FA} (False Alarm time) — суммарная длительность временны́х интервалов, в которых система ошибочно детектирует наличие речи при ее фактическом отсутствии, T_{MISS} (Missed Speech time) — суммарная длительность интервалов, в которых присутствует речь в эталонной разметке, но система не обнаруживает ее, T_{ERR} (Speaker Error time) — суммарная длительность интервалов, в которых речь корректно обнаружена, однако говорящий идентифицирован неверно, T_{TOTAL} — общая длительность речи согласно эталонной (референсной) разметке.

Важно отметить, что вклад в DER вносят три различных типа ошибок, отражающих разные аспекты качества системы. Компонент T_{FA} характеризует избыточную чувствительность детектора речи. Величина T_{MISS} отражает неспособность системы корректно обнаруживать речевую активность. Компонент T_{ERR} оценивает точность разделения и идентификации говорящих при условии корректного обнаружения речи.

Пусть $S = \{A, B, C\}$ — множество спикеров на записи, а $N = \{1, 2, 3\}$ — полученные кластеры из системы диаризации. На рис. 1 показана эталонная разметка трех спикеров с перекрытием между ними. Видно, что в интервалах 30–35 с и 65–70 с одновременно активны два спикера, что учитывается при расчете DER. Результат работы системы диаризации показан на рис. 2. Как видно из пересечения с разметкой, каждый кластер может содержать речь нескольких спикеров.

Для корректного вычисления T_{ERR} необходимо сопоставить кластеры, полученные системой, с реальными спикерами. Это достигается с помощью венгерского алгоритма (Hungarian algorithm [16]), минимизирующего суммарную ошибку идентификации. Матрица стоимостей C_{ij} определяется как разность между общей длительностью кластера и временем пересечения с каждым спикером:

$$C_{ij} = T^{(i)}_{\mathrm{cluster}} - T_{ij},$$

где $T^{(i)}_{\mathrm{cluster}}$ — длительность i-го кластера, T_{ij} — время пересечения кластера i со спикером j (табл. 1).

Матрица стоимостей, используемая для алгоритма, приведена в табл. 2. Оптимальное сопоставление имеет вид: Cluster 1 → Speaker A, Cluster 2 → Speaker B, Cluster 3 → Speaker C.

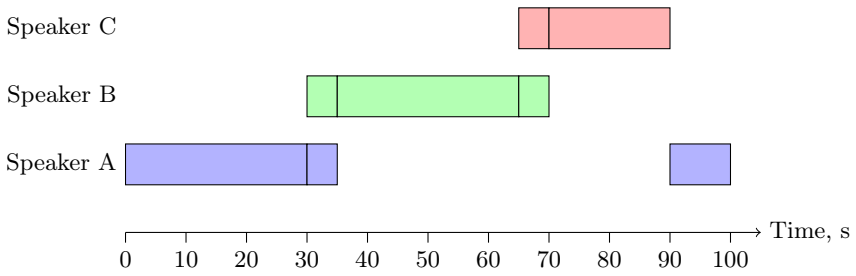

Рис. 1. Известная разметка спикеров

Fig. 1. Ground-truth speaker annotation

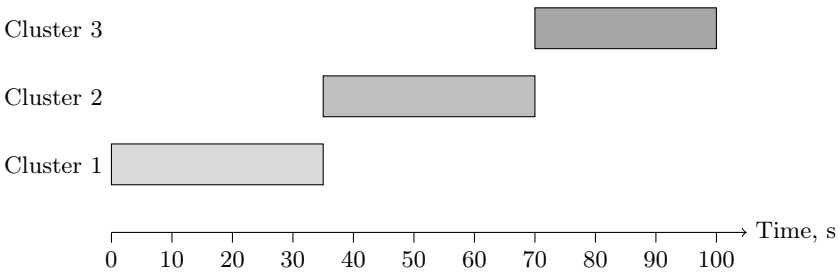

Рис. 2. Прогноз кластеров от системы диаризации

Fig. 2. Cluster prediction from the diarization system

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

Таблица 1. Перекрытия между разметкой и кластерами (в секундах)

Table 1. Overlaps between annotation and clusters (in seconds)

Cluster	Speaker A	Speaker B	Speaker C	Total
1 (0–35)	35	5	0	40
2 (35–70)	0	35	5	40
3 (70–100)	10	0	20	30

Таблица 2. Матрица стоимостей для венгерского алгоритма [16] (в секундах)

Table 2. Cost matrix for the Hungarian algorithm [16] (in seconds)

	Speaker A	Speaker B	Speaker C
Cluster 1	**5**	35	40
Cluster 2	40	**5**	35
Cluster 3	20	30	**10**

После выполнения процедуры сопоставления ошибка идентификации T_{ERR} вычисляется как суммарное время, в течение которого внутри кластера речь принадлежит другому спикеру:

$$T_{\text{ERR}} = 5 + 5 + 10 = 20 \text{ с.}$$

При отсутствии ложных срабатываний ($T_{\text{FA}} = 0$) и пропусков речи ($T_{\text{MISS}} = 0$) DER составит:

$$\text{DER} = \frac{T_{\text{ERR}}}{T_{\text{TOTAL}}} \cdot 100\% = \frac{20}{110} \cdot 100\% \approx 18.2\%.$$

4. Предлагаемый русскоязычный тестовый набор данных RusAudioForum. Разработка и оценка алгоритмов диаризации спикеров требует наличия репрезентативных наборов данных, отражающих характеристики целевых приложений. Существующие стандартные наборы данных (AMI [4], VoxConverse [5]), хотя и обеспечивают сравнительную оценку методов, имеют ряд ограничений: большинство из них созданы на английском языке, что затрудняет оценку методов для русскоязычных приложений; многие датасеты включают записи ограниченной продолжительности (до 10–15 мин), недостаточные для оценки устойчивости алгоритмов при длительных записях; кроме того, существующие наборы данных часто не отражают специфику сценариев публичных мероприятий — форумов, конференций и панельных дискуссий, которые характеризуются большой продолжительностью (60 мин и более), переменным числом спикеров (от 2 до 8 и более), естественными условиями многоспикерного взаимодействия и специфическими акустическими условиями.

Отметим, что набор данных DIHARD-III [17], который широко использовался для оценки методов диаризации, в настоящее время стал недоступен для скачивания, поэтому результаты по нему в данной работе не представлены.

Создание нового русскоязычного набора данных для домена публичных мероприятий необходимо для решения нескольких задач. Отсутствие публичных русскоязычных наборов данных для диаризации затрудняет развитие и оценку методов, специфичных для русского языка, что особенно важно, учитывая фонетические и просодические особенности русской речи. Кроме того, домен публичных мероприятий представляет особый интерес для практических приложений, где требуется обработка длительных многоспикерных записей в режиме реального времени, что обусловливает необходимость оценки устойчивости алгоритмов на записях продолжительностью 60–200 мин.

Для экспериментов с многоспикерными записями был создан набор данных RusAudioForum. Он включает записи известных публичных мероприятий, таких как Санкт-Петербургский международный экономический форум (ПМЭФ), различные конференции, форумы и панельные дискуссии. Эти события характеризуются высоким качеством аудиозаписи, разнообразием спикеров и естественными условиями многоспикерного взаимодействия, что делает их репрезентативными для задачи онлайн-диаризации в реальных сценариях.

4.1. Разметка тестового набора данных. Набор создан на основе данных, полученных с использованием сервиса TagMe от Сбербанка [18] — системы автоматической разметки и аннотирования муль-

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

тимедийного контента. Этот сервис представляет собой платформу для разметки различных форматов данных с многоуровневой системой валидации результатов разметки.

Разметка набора данных осуществлялась в несколько этапов. Изначально аудиозаписи публичных мероприятий обрабатывались через сервис TagMe для получения предварительной разметки спикеров. Затем проводилась ручная верификация и корректировка временны́х меток специалистами, что обеспечило высокое качество эталонной разметки. Для каждого сегмента речи были определены точные временны́е границы начала и окончания высказывания, а также анонимизированный идентификатор соответствующего спикера.

4.2. Основные характеристики тестового набора данных. Набор данных характеризуется разнообразием акустических условий, длительностью записей и количеством участвующих спикеров.

В табл. 3 представлены ключевые статистические характеристики набора данных, отражающие его структуру и сложность. Всего в наборе 18 файлов. В каждом файле реплики размечены в формате “номер спикера — время начала реплики — время окончания реплики”.

Представленные характеристики подтверждают репрезентативность набора данных для задачи онлайн-диаризации в реальных условиях публичных мероприятий, где требуется обработка длительных записей с переменным числом спикеров и значительным количеством перекрытий речи.

5. Построение системы онлайн-диаризации. В рамках данной работы предложена система по обработке речевых сигналов в онлайн режиме с автоматическим распознаванием речи и диаризацией спикеров. Общая схема представлена на рис. 3. Основными компонентами являются легковесный и информативный эмбеддер ResNet34 и алгоритм дискриминантного анализа, реализованный при помощи подхода на основе PSDA.

5.1. Вероятностное сравнение спикерских эмбеддингов. В качестве основного метода вероятностного моделирования спикерских эмбеддингов в нашей системе используется Probabilistic Spherical Discriminant Analysis (PSDA) [19], который было предложено адаптировать для потокового распознавания речи. Данный подход специально создан для работы с нормированными эмбеддингами, лежащими на единичной гиперсфере $\mathbb{S}^{d-1}$ (алгоритм 1).

Идея PSDA состоит в построении генеративной вероятностной модели, описывающей, как наблюдаемые эмбеддинги $\boldsymbol{e} \in \mathbb{S}^{d-1}$ порождаются скрытыми факторами спикеров. Для каждого спикера s вводится латентный вектор $\boldsymbol{y}_s$ в низкоразмерном дискриминантном подпространстве и предполагается, что отдельные эмбеддинги $\boldsymbol{e}_{s,i}$ этого спикера распределены на сфере вокруг направления, определяемого $\boldsymbol{y}_s$,

Таблица 3. Характеристики датасета

Table 3. Dataset characteristics

Параметр Parameter	Значение Value
Количество файлов Number of files	18
Общее количество часов Total number of hours	~ 20
Количество уникальных говорящих для одного файла Number of unique speakers per file	от 3 до 11 from 3 to 11
Количество реплик спикеров для одного файла Number of speaker replicas per file	от 14 до 208 from 14 to 208
Средняя длительность реплик спикеров для одного файла Mean duration of speaker replicas per file	от 15.8 с до 229 с from 15.8 s to 229 s
Максимальная длительность дискуссии Maximum duration of discussion	~ 200 с ~ 200 s
Средняя длительность дискуссии Average duration of discussion	~ 100 с ~ 100 s
Перекрытие речи Speech overlap	$0.0\% - 3.5\%$

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

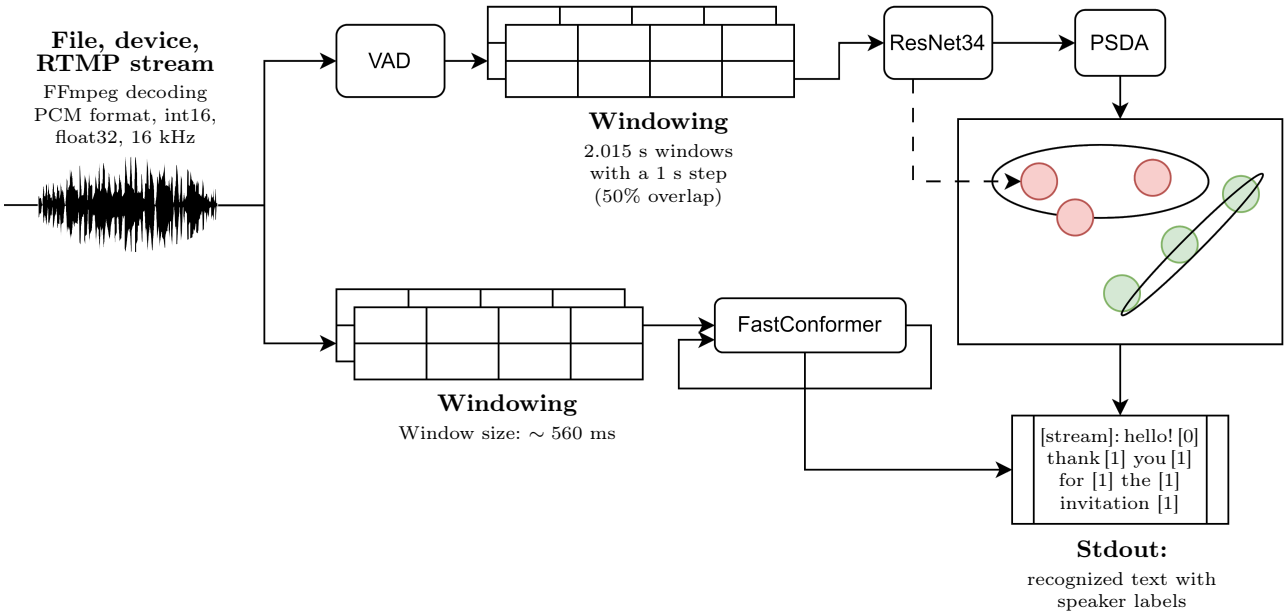

Рис. 3. Архитектура системы онлайн-диаризации спикеров с модулем автоматического распознавания речи (с программной реализацией системы можно ознакомиться, перейдя по ссылке https://github.com/pianonyy/OnlineDIASR)

Fig. 3. Architecture of the online speaker diarization pipeline with automatic speech recognition module (the software implementation of the system is available at https://github.com/pianonyy/OnlineDIASR)

Алгоритм 1. Онлайн-диаризация спикеров на основе PSDA

Algorithm 1. Online speaker diarization based on PSDA

require: Audio stream A
  while new audio frames arrive **do**
    $V_t := \mathrm{VAD}(A_t)$
    for each speech segment $v \in V_t$ **do**
      $x := \text{log-Mel}(v)$
      $e := \mathrm{EmbeddingModel}(x)$
      optionally: $e := \mathrm{PCA}(e)$
      $s := \mathrm{PSDA\_Cluster}(e)$
      save $\{start, end, s\}$
    end for
  end while
  merge adjacent segments with identical s

с концентрационным параметром κ_{w}, отвечающим за внутриспикерскую вариативность. Межспикерские различия моделируются через распределение для $\boldsymbol{y}_s$ с собственным параметром концентрации κ_{b}, регулирующим степень разделимости спикеров.

Формально плотность вероятности распределения эмбеддингов спикера, заданного латентным вектором $\boldsymbol{y}$, задается распределением фон Мизеса–Фишера:

$$p(\boldsymbol{e}_{s,i} \mid \boldsymbol{y}_s) \propto \exp\left(\kappa_{\mathrm{w}}\, \boldsymbol{e}_{s,i}^T \frac{W\boldsymbol{y}_s}{\|W\boldsymbol{y}_s\|}\right),$$

где W — матрица проекции в дискриминантное подпространство. Аналогично задается априорное распределение для $\boldsymbol{y}_s$ с параметром κ_{b}. Такое построение обеспечивает естественную интерпретацию: чем

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

ближе эмбеддинги к направлению кластера на сфере, тем более правдоподобна гипотеза о том, что они принадлежат одному и тому же спикеру.

Практически PSDA обучается по принципу максимального правдоподобия с использованием пар или наборов эмбеддингов, помеченных по говорящему, что позволяет оценить матрицу W и параметры концентрации κ_{w}, κ_{b}. После обучения модель используется для вычисления логарифма отношения правдоподобий (log-likelihood ratio, LLR) между гипотезами H_{same} (эмбеддинги принадлежат одному спикеру) и H_{diff} (разным спикерам). Эти LLR-оценки служат основой для принятия решений в кластеризаторе (раздел 5.2).

В онлайн-сценарии PSDA естественным образом интегрируется в потоковую обработку речи: по мере поступления новых эмбеддингов для каждого активного кластера пересчитываются апостериорные оценки спикерских факторов $\boldsymbol{y}_s$ и соответствующие LLR-значения. Это позволяет, во-первых, инкрементально уточнять параметры кластеров без пересчета по всей истории; во-вторых, вводить пороги для создания новых спикеров в сценарии с их неограниченным количеством; и в-третьих, стабилизировать решения за счет байесовского сглаживания правдоподобий.

5.2. Общая архитектура. В состав разработанной системы входят два ключевых компонента: потоковый модуль распознавания речи (на базе [3]) и модуль онлайн-диаризации спикеров.

Для извлечения эмбеддингов спикеров используется архитектура ResNet-34 — глубокая сверточная нейронная сеть, состоящая из 34 слоев, адаптированная для обработки речевых сигналов. Архитектура ResNet (Residual Network) основана на использовании остаточных (residual) связей, которые позволяют эффективно обучать глубокие сети, предотвращая проблему затухания градиентов при обратном распространении ошибки.

Выбор ResNet-34 для задачи извлечения эмбеддингов спикеров обусловлен несколькими преимуществами данной архитектуры. Во-первых, сверточные слои ResNet эффективно извлекают признаки из мел-спектрограмм речевого сигнала, что критично для различения индивидуальных характеристик голосов спикеров. Во-вторых, архитектура обеспечивает баланс между глубиной и вычислительной эффективностью, что важно для онлайн-диаризации, где требуется быстрая обработка в реальном времени. В-третьих, ResNet-34 показала высокую эффективность в задачах верификации и идентификации спикеров, демонстрируя конкурентоспособные результаты на стандартных наборах данных, таких как VoxCeleb [20]. По сравнению с моделями большей глубины (ResNet-50, ResNet-101), ResNet-34 требует меньше вычислительных ресурсов при сохранении качества извлечения эмбеддингов, что делает ее оптимальным выбором для потоковой обработки.

Модель ResNet-34 была обучена с помощью библиотеки WeSpeaker [21]. Эта библиотека предоставляет комплексный набор инструментов для обучения моделей верификации и идентификации спикеров, включая реализацию современных архитектур (ResNet, ECAPA–TDNN) и методов оценки качества эмбеддингов. Библиотека поддерживает обучение на крупных датасетах (таких как VoxCeleb [20]) и предоставляет предобученные модели, оптимизированные для извлечения высококачественных эмбеддингов спикеров. Использование предобученной модели из WeSpeaker позволяет применять проверенные на практике конфигурации обучения.

Для предотвращения ложных переключений между спикерами применяются три простые эвристики:

1) запрет смены спикера на фрагментах короче 1 с;
2) разрешение смены при паузах длиной более 1.5 с;
3) перекрытие окон для плотного покрытия речевого сигнала.

5.3. Датасеты. Для комплексной оценки предложенного метода использованы как стандартные публичные наборы данных, так и собственный русскоязычный набор данных.

Стандартные наборы данных:

1. AMI [4] — корпус записей совещаний с известной разметкой спикеров, широко используемый для оценки методов диаризации;
2. VoxConverse [5] — набор данных из интервью и разговоров, характеризующийся естественными условиями записи.

Русскоязычный тестовый набор данных RusAudioForum представляет собой набор данных (∼20 часов), включающий записи публичных событий: конференций, форумов, панельных дискуссий. Набор данных опубликован на `Hugging Face`, а реализация программных модулей для диаризации и распознавания речи — на `GitHub`.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

6. Эксперименты.

6.1. Методы сравнения. Для оценки предложенного метода проведено сравнение со следующими подходами:

1. Sortformer [2] — end-to-end метод онлайн-диаризации (используется только для записей с числом спикеров, не превышающим 4);
2. Pyannote.audio 2.1 и 3.1 [13] — офлайн-методы для референсного сравнения.

В экспериментах дополнительно оценивается вариант, в котором перед PSDA применяется снижение размерности эмбеддингов методом главных компонент (PCA). Исходные эмбеддинги ResNet-34 имеют размерность 256. После L2-нормализации к ним применяется предобученная PCA-модель, сокращающая размерность до 128. Таким образом, цепочка обработки включает следующие этапы: извлечение эмбеддинга, нормализация, PCA, моделирование PSDA. Для 128-мерных эмбеддингов используется отдельная PSDA-модель, обученная в пространстве сниженной размерности.

Снижение размерности приводит к уменьшению объема памяти на эмбеддинг (что может быть критично при большом количестве спикеров) и ускорению вычислений при кластеризации и подсчете LLR. При этом главные компоненты сохраняют основную долю дисперсии, связанную с межспикерными различиями, благодаря чему качество диаризации остается высоким.

Для набора данных RusAudioForum модель Softformer прогонялась на 4 файлах из датасета (из-за ограничения на 4 спикеров у Softformer). По результатам экспериментов (табл. 4) можно сделать следующие выводы:

1. Предложенный подход онлайн-диаризации снижает значение метрики DER в 3 раза (с 25.73% до 6.81%) по сравнению с моделью Nvidia Softformer (2025), тем самым показывая лучше результаты на целевом датасете RusAudioForum. Эвристики, описанные в разделе 5, улучшили качество распознавания спикеров.
2. Датасет AMI является наиболее сложным для всех систем. Наблюдаются самые высокие значения DER, достигающие 36.47% (по сравнению с VoxConverse и RusAudioForum). Это связано с тем, что AMI содержит аудиозаписи более низкого качества.
3. Популярный инструмент Pyannote показывает хорошие результаты, сравнимые с предложенным подходом. Однако модель Pyannote не может быть применена для потоковой онлайн обработки аудиосигнала.

6.2. Анализ ошибок онлайн-диаризации. Для качественной оценки поведения систем рассмотрен фрагмент записи длительностью 90 с (интервал 80.3–170.3 с). На рис. 4–6 приведены сегменты разметки в формате (время начала, длительность, ID спикера) для эталонной разметки (GT), предложенного подхода и модели Softformer для 1 примера из набора RusAudioForum. Эталон содержит два длинных сегмента (два спикера); предложенный подход дает три сегмента с одним кратким переключением; Softformer

Таблица 4. Сравнение систем по метрике DER (в %) на стандартных наборах данных и предложенном датасете RusAudioForum (чем меньше DER, тем лучше). Для Sortformer на RusAudioForum указаны результаты только для записей с не более чем 4 спикерами. Сокращение $\pm$*heur.* обозначает добавление/удаление эвристик, предложенных в разделе 5

Table 4. Comparison of systems by DER metric (in %) on standard benchmarks and the proposed RusAudioForum dataset (lower DER is better). For Sortformer results on RusAudioForum are reported only for recordings with no more than 4 speakers. The term $\pm$*heur.* denotes the addition/removal of heuristics proposed in Section 5

Модель Model	Режим обработки Processing mode	AMI	VoxConverse		RusAudioForum	
Pyannote 2.1	offline	27.12	11.24		10.93	
Pyannote 3.1	offline	**22.4**	11.3		**6.40**	
Sortformer v1	online	30.34	16.76		25.73	
			+*heur.*	−*heur.*	+*heur.*	−*heur.*
ResNet34 + CSEA (Ours)	online	29.67	**11.04**	11.29	7.57	8.09
ResNet34 + PSDA (Ours)	online	36.47	14.55	14.96	8.85	12.69
ResNet34 + PCA + PSDA (Ours)	online	27.82	17.40	18.58	6.81	9.22

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

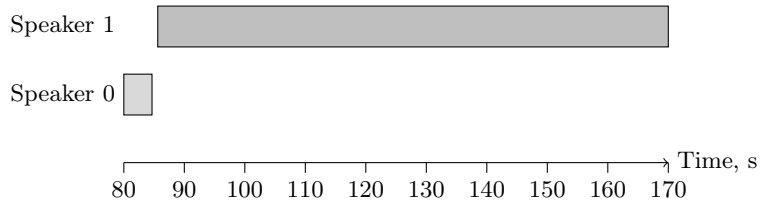

Рис. 4. Эталонная разметка (GT) для фрагмента [80.3–170.3 с]

Fig. 4. Ground-truth annotation (GT) for the fragment [80.3–170.3 s]

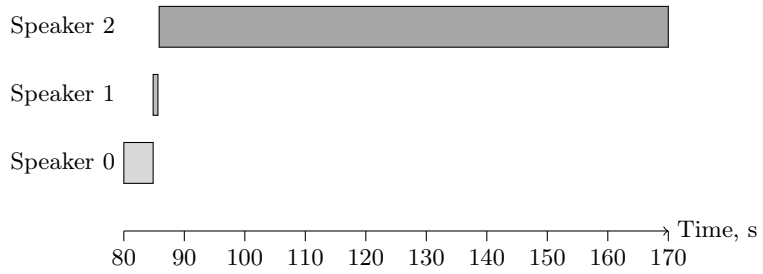

Рис. 5. Разметка предложенного подхода для фрагмента [80.3–170.3 с]

Fig. 5. Annotation of the proposed pipeline for the fragment [80.3–170.3 s]

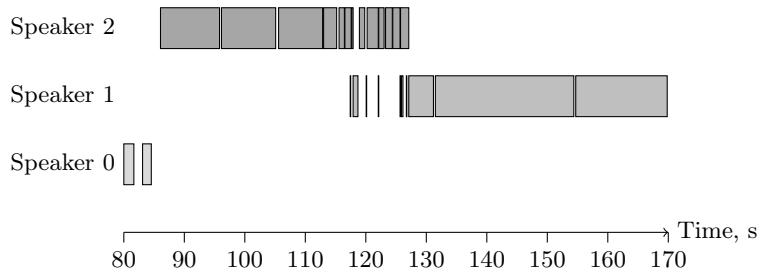

Рис. 6. Разметка Softformer для фрагмента [80.3–170.3 с]: множество коротких переключений между спикерами

Fig. 6. Softformer annotation for the fragment [80.3–170.3 s]: many short switches between speakers

порождает множество коротких переключений между спикерами. В данном окне предложенный подход ближе к GT по метрике DER за счет меньшего числа ложных переключений спикеров.

7. Заключение. В данной работе представлен подход потокового распознавания речи с модулем онлайн-диаризации спикеров, поддерживающий неограниченное число спикеров. Подходы онлайн-диаризации важны для приложений потоковой обработки, в особенности для систем синхронного перевода в реальном времени. Алгоритм включает модульную каскадную архитектуру (VAD, ResNet-эмбеддинги, PSDA), а также эвристики, обеспечивающие устойчивость к коротким паузам и шуму.

Перечислим основные результаты настоящей работы:

1. Подготовлен и опубликован тестовый набор данных RusAudioForum для задачи онлайн-диаризации спикеров с разметкой. Набор данных включает записи известных публичных мероприятий. Эти события характеризуются высоким качеством аудиозаписи, разнообразием спикеров и естественными условиями многоспикерного взаимодействия.
2. Предложенный подход онлайн-диаризации представляет собой каскадную обработку речи, основанную на применении ResNet-эмбеддингов, методов PSDA и эвристик устойчивости.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

3. На созданном целевом датасете RusAudioForum предложенный подход снижает значение метрики DER в 3 раза (с 25.73% до 6.81%) по сравнению со сквозной моделью Nvidia Softformer (2025), тем самым показывая лучшие результаты. Эвристики, предложенные в разделе 5, улучшили качество распознавания спикеров.
4. Популярный инструмент Pyannote показывает хорошие результаты, сравнимые с предложенным подходом. Однако модель Pyannote не может быть применена для потоковой онлайн обработки аудиосигнала.

В будущем планируется подробнее исследовать адаптивный подбор параметров для вероятностного моделирования эмбеддингов спикеров в различных акустических сценариях.

Список литературы

1. *Liang D., Li X.* LS-EEND: long-form streaming end-to-end neural diarization with online attractor extraction // IEEE Trans. Audio Speech Lang. Process. 2025. **33**. 3568–3581. doi 10.1109/TASLPRO.2025.3597446.
2. *Park T., Medennikov I., Dhawan K., et al.* Sortformer: seamless integration of speaker diarization and ASR by bridging timestamps and tokens // arXiv preprint arXiv:2409.06656, 2024. doi 10.48550/arXiv.2409.06656.
3. *Noroozi V., Majumdar S., Kumar A., et al.* Stateful conformer with cache-based inference for streaming automatic speech recognition // 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 2024. IEEE Press, 2024. pp. 12041–12045. doi 10.1109/ICASSP48485.2024.10446861.
4. *Carletta J., Ashby S., Bourban S., et al.* The AMI meeting corpus: a pre-announcement // Int. Workshop on Machine Learning for Multimodal Interaction, 2005. Lecture Notes in Computer Science, Vol. 3869. Berlin: Springer, 2006. pp. 28–39. doi 10.1007/11677482_3.
5. *Chung J.S., Huh J., Nagrani A., et al.* Spot the conversation: speaker diarisation in the wild // Proc. 21th Interspeech Conference, Shanghai, October 25–29, 2020. Interspeech Press, 2020. pp. 299-303. doi 10.21437/Interspeech.2020-2337.
6. *Aperdannier R., Schacht S., Piazza A.* A review of common online speaker diarization methods // arXiv preprint arXiv:2406.14464, 2024. doi 10.48550/arXiv.2406.14464.
7. *Dehak N., Kenny P.J., Dehak R., et al.* Front-end factor analysis for speaker verification // IEEE Trans. Audio Speech Lang. Process. 2011. **19**, N 4. 788–798. doi 10.1109/TASL.2010.2064307.
8. *Sell G., Garcia-Romero D.* Speaker diarization with PLDA i-vector scoring and unsupervised calibration // 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, December 07–10, 2014. IEEE Press, 2014. pp. 413–417. doi 10.1109/SLT.2014.7078610.
9. *Variani E., Lei X., McDermott E., et al.* Deep neural networks for small footprint text-dependent speaker verification // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 04–09, 2014. IEEE Press, 2014. pp. 4052–4056. doi 10.1109/ICASSP.2014.6854363.
10. *Desplanques B., Thienpondt J., Demuynck K.* ECAPA–TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification // Proc. 21th Interspeech Conference, Shanghai, October 25–29, 2020. Interspeech Press, 2020. pp. 3830–3834. doi 10.21437/Interspeech.2020-2650.
11. *Wan L., Wang Q., Papir A., Moreno I.L.* Generalized end-to-end loss for speaker verification // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, April 15–20, 2018. IEEE Press, 2018. pp. 4879–4883. doi 10.1109/ICASSP.2018.8462665.
12. *Sholokhov A., Kuzmin N., Lee K.A., Chng E.S.* Probabilistic back-ends for online speaker recognition and clustering // 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June 04–10, 2023. IEEE Press, 2023. pp. 1–5. doi 10.1109/ICASSP49357.2023.10097032.
13. *Bredin H., Yin R., Coria J.M., et al.* Pyannote.audio: neural building blocks for speaker diarization // ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, May 04–08, 2020. IEEE Press, 2020. pp. 7124–7128. doi 10.1109/ICASSP40776.2020.9052974.
14. *Zhang A., Wang Q., Zhu Z., et al.* Fully supervised speaker diarization // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12–17, 2019. IEEE Press, 2019. pp. 6301–6305. doi 10.1109/ICASSP.2019.8683892.
15. *Coria J.M., Bredin H., Ghannay S., Rosset S.* Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation // 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, December 13–17, 2021. IEEE Press, 2021. pp. 1139–1146. doi 10.1109/ASRU51503.2021.9688044.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

16. *Kuhn H.W.* The Hungarian method for the assignment problem // Naval Research Logistics Quarterly. 1955. **2**, N (1–2). 83–97. doi 10.1002/nav.3800020109.
17. *Ryant N., Singh P., Krishnamohan V., et al.* The Third DIHARD diarization challenge // 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czechia, August 30–September 3, 2021. Interspeech Press, 2021. pp. 3570–3574. doi 10.21437/Interspeech.2021-1208.
18. TagMe – Сервис автоматической разметки мультимедиа. https://developers.sber.ru/portal/products/tagme. Дата обращения: 5 апреля 2026.
19. *Brümmer N., Swart A., Mošner L., et al.* Probabilistic spherical discriminant analysis: an alternative to PLDA for length-normalized embeddings // arXiv preprint arXiv:2203.14893, 2022. doi 10.48550/arXiv.2203.14893.
20. *Nagrani A., Chung J.S., Zisserman A.* VoxCeleb: a large-scale speaker identification dataset // Proceedings Interspeech, Stockholm, Sweden, August 20—24, 2017. Interspeech Press, 2017. pp. 2616–2620. doi 10.21437/Interspeech.2017-950.
21. *Wang H., Liang C., Wang S., et al.* Wespeaker: A research and production oriented speaker embedding learning toolkit // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June 4–10, 2023. IEEE Press, 2023. pp. 1–5. doi 10.1109/ICASSP49357.2023.10096626.

Получена 17 февраля 2026 г. | *Принята* 30 марта 2026 г. | *Опубликована* 24 апреля 2026 г.

Информация об авторах

Антон Вячеславович Полевой — аспирант; Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, 1, стр. 4, 119234, Москва, Российская Федерация.

Наталья Валентиновна Лукашевич — д.т.н., профессор; Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Научно-исследовательский вычислительный центр, Ленинские горы, 1, стр. 4, 119234, Москва, Российская Федерация.

References

1. D. Liang and X. Li, “LS-EEND: Long-Form Streaming End-to-End Neural Diarization with Online Attractor Extraction,” IEEE Trans. Audio Speech Lang. Process. **33**, 3568–3581 (2025). doi 10.1109/TASLPRO.2025.3597446.
2. T. Park, I. Medennikov, K. Dhawan, et al., “Sortformer: Seamless Integration of Speaker Diarization and ASR by Bridging Timestamps and Tokens,” arXiv preprint arXiv:2409.06656, 2024. doi 10.48550/arXiv.2409.06656.
3. V. Noroozi, S. Majumdar, A. Kumar, et al., “Stateful Conformer with Cache-Based Inference for Streaming Automatic Speech Recognition,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 2024* (IEEE Press, 2024), pp. 12041–12045. doi 10.1109/ICASSP48485.2024.10446861.
4. J. Carletta, S. Ashby, S. Bourban, et al., “The AMI Meeting Corpus: A Pre-announcement,” in *Int. Workshop on Machine Learning for Multimodal Interaction, 2005* Lecture Notes in Computer Science, Vol. 3869. (Springer, Berlin, 2006), pp. 28–39. doi 10.1007/11677482_3.
5. J. S. Chung, J. Huh, A. Nagrani, et al., “Spot the Conversation: Speaker Diarisation in the Wild,” in *Proc. 21th Interspeech Conference, Shanghai, October 25–29, 2020.* (Interspeech Press, 2020), pp. 299-303. doi 10.21437/Interspeech.2020-2337.
6. R. Aperdannier, S. Schacht, and A. Piazza, “A Review of Common Online Speaker Diarization Methods,” arXiv preprint arXiv:2406.14464, 2024. doi 10.48550/arXiv.2406.14464.
7. N. Dehak, P. J. Kenny, R. Dehak, et al., “Front-End Factor Analysis for Speaker Verification,” IEEE Transactions on Audio, Speech, and Language Processing **19** (4), 788–798 (2011). doi 10.1109/TASL.2010.2064307.
8. G. Sell and D. Garcia-Romero, “Speaker diarization with PLDA i-vector scoring and unsupervised calibration,” in *2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, December 07–10, 2014* (IEEE Press, 2014), pp. 413–417. doi 10.1109/SLT.2014.7078610.
9. E. Variani, X. Lei, E. McDermott, et al., “Deep neural networks for small footprint text-dependent speaker verification,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, May 04–09, 2014* (IEEE Press, 2014), pp. 4052–4056. doi 10.1109/ICASSP.2014.6854363.

2026, **27** (2), 194–207. doi 10.26089/NumMet.v27r213

10. B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA–TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. 21th Interspeech Conference, Shanghai, October 25–29, 2020.* (Interspeech Press, 2020), pp. 3830–3834. doi 10.21437/Interspeech.2020-2650.
11. L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, April 15–20, 2018.* (IEEE Press, 2018), pp. 4879–4883. doi 10.1109/ICASSP.2018.8462665.
12. A. Sholokhov, N. Kuzmin, K. A. Lee, and E. S. Chng, "Probabilistic Back-Ends for Online Speaker Recognition and Clustering," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June 04–10, 2023.* (IEEE Press, 2023), pp. 1–5. doi 10.1109/ICASSP49357.2023.10097032.
13. H. Bredin, R. Yin, J. M. Coria, et al., "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, May 04–08, 2020, (IEEE Press, 2020), pp. 7124–7128. doi 10.1109/ICASSP40776.2020.9052974.
14. A. Zhang, Q. Wang, Z. Zhu, et al., "Fully Supervised Speaker Diarization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, May 12–17, 2019.* (IEEE Press, 2019), pp. 6301–6305. doi 10.1109/ICASSP.2019.8683892.
15. J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Overlap-Aware Low-Latency Online Speaker Diarization Based on End-to-End Local Segmentation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, December 13–17, 2021.* (IEEE Press, 2021), pp. 1139–1146. doi 10.1109/ASRU51503.2021.9688044.
16. H. W. Kuhn, "The Hungarian method for the assignment problem," Naval Research Logistics Quarterly **2** (1–2), 83–97 (1955). doi 10.1002/nav.3800020109.
17. N. Ryant, P. Singh, V. Krishnamohan, et al., "The Third DIHARD Diarization Challenge," in *22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021), Brno, Czechia, August 30–September 3, 2021.* (Interspeech Press, 2021), pp. 3570–3574. doi 10.21437/Interspeech.2021-1208.
18. Automatic multimedia annotation service TagMe. https://developers.sber.ru/portal/products/tagme. Cited April 5, 2026.
19. N. Brümmer, A. Swart, L. Mošner, et al., "Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings," arXiv preprint arXiv:2203.14893, 2022. doi 10.48550/arXiv.2203.14893.
20. A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proceedings Interspeech, Stockholm, Sweden, August 20—24, 2017.* (Interspeech Press, 2017), pp. 2616–2620. doi 10.21437/Interspeech.2017-950.
21. H. Wang, C. Liang, S. Wang, et al., "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, June 4–10, 2023.* (IEEE Press, 2023), pp. 1–5. doi 10.1109/ICASSP49357.2023.10096626.

Received	*Accepted*	*Published*
February 17, 2026	*March 30, 2026*	*April 24, 2026*

Information about the authors

Anton A. Polevoi — PhD student; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Leninskie Gory, 1, building 4, 119234, Moscow, Russia.

Natalya V. Loukachevitch — Dr. Sci., Professor; Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Research Computing Center, Leninskie Gory, 1, building 4, 119234, Moscow, Russia.