УДК 004.934.2
DOI:10.25729/ESI.2026.43.3.003

Адаптивная гибридная нейросетевая архитектура на основе внимания для идентификации дикторов в корпоративных совещаниях

Потапов Арсений Денисович[1,2], Лукьянов Никита Дмитриевич[2]

[1]ПАО Сбербанк России, Россия, Иркутск
[2]Иркутский национальный исследовательский технический университет,
Россия, Иркутск, *lukyanovnd@istu.edu*

Аннотация. Актуальность исследования обусловлена ростом потребности в автоматизации ведения протоколов корпоративных встреч, которое является трудоёмким и затяжным процессом, сопровождающимся высоким риском ошибок. Современные системы распознавания речи (ASR – Automatic Speech Recognition) работают нестабильно при многопользовательских диалогах, перекрытиях голосовых фрагментов, фоновом шуме и использовании специализированной лексики. Особенно остро стоит проблема распознавания дикторов (Speaker Identification, SID), когда на одного сотрудника приходится мало эталонных записей, что вполне типично для офисной среды. Дополнительные трудности создаёт и акустическая обстановка: гулкие кабинеты, одновременные высказывания, отсутствие чёткого разделения голосов. Таким образом, цель данной работы – создание гибридной нейросетевой модели для точной транскрипции и надёжной атрибуции речи в групповых дискуссиях. В рамках работы проанализированы слабые места ASR-систем и методов SID, предложена трёхуровневая архитектура: сегментация аудиопотока VAD + SCD (VAD – Voice Activity Detection – определение активности речи), SCD – Speaker Change Detection – детектирование смены диктора), идентификация по голосовым эмбеддингам с механизмом внимания к эталонам, транскрибация[1] с учётом личности говорящего, так же приведено математическое обоснование ключевых компонентов. Теоретическая значимость заключается в построении методики SID на основе механизма внимания, позволяющая эффективно решать задачи обучения с малым числом примеров (few-shot learning). В отличие от традиционных подходов, требующих обширных размеченных корпусов, новая архитектура учится сопоставлять речевые фрагменты с небольшим числом эталонных записей. Эффективность метода эмпирически подтверждена на реальных записях корпоративных совещаний (свыше 20 часов, пять участников). Модуль сегментации показал F1-меру VAD на уровне 0,885, а модуль идентификации диктора точность 62,5%, что в трое выше случайного уровня (20%). Транскрибация выполнена через внешнюю систему Whisper, гарантирующую высокую достоверность распознавания. Результаты позволяют построить систему автоматического ведения протоколов совещаний, а архитектура позволяет работать и с большим числом участников, а также легко адаптируется под разные форматы коммуникаций: от переговоров до встреч с участием клиентов. Функциональные возможности гибридной нейросетевой модели могут быть использованы при создании различных инструментов для протоколирования корпоративных встреч.

Ключевые слова: автоматическое распознавание речи, идентификация диктора, гибридная нейросетевая архитектура, механизм внимания, голосовые эмбеддинги, сегментация аудиопотока, транскрибация, корпоративные совещания

Цитирование: Потапов А.Д. Адаптивная гибридная нейросетевая архитектура на основе внимания для идентификации дикторов в корпоративных совещаниях / А.Д. Потапов, Н.Д. Лукьянов // Информационные и математические технологии в науке и управлении, 2026. – № 3(43). – С. 31-41. – DOI:10.25729/ESI.2026.43.3.003.

Введение. Запись совещаний, презентаций и рабочих обсуждений давно стала привычной практикой в большинстве компаний. Архив таких аудиофайлов – это слепок деловой активности: в нём остаются детали проектных решений, логика управленческих шагов и контекст принятых договорённостей. Однако выделить из этих данных структурированное содержание, кто именно, когда и что произнёс, до сих пор является многочасовым кропотливым трудом [1]. Цифровая трансформация бизнеса заставляет крупные компании искать способы автоматической фиксации всего, что происходит на

[1] Транскрибация – процесс перевода устной речи из аудио или видеозаписи в текстовый формат.

встречах – от еженедельных совещаний команды разработки до многочасовых заседаний архитектурного комитета.

Отметим существенное различие решаемой задачи от задач, решаемых системами автоматического распознавания речи, ориентированных на расшифровку монологов, записанных в звукоизолированной кабине [2]. При использовании ASR в реальных корпоративных дискуссиях возникают ошибки: несколько участников говорят по очереди или одновременно, перебивают друг друга, паузы между репликами минимальны, что в конечном счете приводит к ухудшению точности стандартных ASR-решений [3]. Недостаток таких систем – это неумение надёжно поймать момент передачи слова и безошибочно привязать сказанное к конкретному сотруднику [4].

Идентификация диктора (SID) и разделение аудиопотока по голосам (диаризация) долгое время рассматривались как самостоятельные задачи, решаемые отдельно от распознавания текста [5]. Применительно к записям рабочих встреч эта разобщённость создаёт дополнительные трудности. Классические подходы к SID, опирающиеся на гауссовы смеси с универсальной фоновой моделью и технику i-векторов, строятся на выделении устойчивых акустических признаков и их статистическом описании [6]. Подобные методы плохо справляются с короткими фрагментами речи, требуют внушительных объёмов записей на каждого человека и чувствительны к шумам, которыми обычно наполнены офисные переговорные [7].

Прорыв в этой области обеспечило глубокое обучение. Свёрточные и рекуррентные сети, а затем и трансформерные архитектуры [8] позволили заметно улучшить качество как самой транскрибации – достаточно вспомнить успехи модели Whisper [9], – так и извлечения голосовых эмбеддингов, где стандартом стали x-vector и d-vector [10]. Впрочем, даже передовые системы диаризации, например, TS-VAD [11], по-прежнему привязаны к предварительной кластеризации эмбеддингов и раскрывают свой потенциал лишь при наличии масштабных обучающих выборок.

Для корпоративной среды характерна иная ситуация. Здесь на каждого сотрудника приходится считанное количество эталонных записей, и традиционные методы машинного обучения в таких условиях теряют точность [12]. Подобное положение дел вынуждает обращаться к архитектурам, рассчитанным на работу при острой нехватке примеров, то есть в режиме few-shot learning. Соответственно, требуется построить единую модель, которая совмещала бы транскрибацию и идентификацию говорящего, причём делала это в условиях ограниченных обучающих данных, столь привычных для офисной среды. Оценку такой системы естественно вести по трём направлениям: насколько точно реплика приписывается автору (Speaker Accuracy), как уверенно вычленяются осмысленные речевые отрезки из шумов и пауз (VAD F1) и насколько безошибочно фиксируется переход слова от одного участника к другому (SCD Precision).

В рамках работы предлагается многомодульная нейросетевая архитектура, берущая на себя одновременно задачи сегментации звукового потока, расшифровки сказанного и установления личности каждого говорящего. Достигнутые на сегодня результаты могут стать основой для разработки средств автоматической фиксации совещаний – инструментов, которые позволят упорядочить документирование корпоративных встреч и снять часть рутинной нагрузки с сотрудников.

Материалы и методы. В основе предлагаемого метода лежит архитектура, объединяющая модули сегментации на основе свёрточной сети, идентификации диктора с механизмом внимания и транскрибации на основе трансформера, предназначенная для последовательной обработки аудиосигнала многопользовательского совещания. Общая схема метода включает три ключевых модуля, работающих последовательно: модуль сегментации и

выделения признаков, модуль идентификации диктора и модуль транскрибации. Обучение сети проводится в два этапа: предварительное обучение на больших публичных наборах данных и тонкая настройка на целевых голосах сотрудников.

Пусть исходными данными является аудиозапись совещания X, представленная в виде временного ряда. Первоначально выполняются предобработка и выделение признаков. Входной сигнал $X(t)$ преобразуется в последовательность логарифмических мел-спектрограмм [13] $S = (s_1, s_2, \dots, s_T)$, где $s_t \in R^F$ – вектор из F мел-частотных коэффициентов в момент времени t, где T – длительность записи [14]. Данное представление сохраняет как спектральные, так и временные характеристики сигнала и является стандартным для современных моделей распознавания речи.

Первым модулем является «Сеть сегментации и выделения универсальных речевых признаков», который реализован на основе стандартной сверточной нейронной сети (CNN) с одномерными свертками во временной области [15]. Его задача – это анализ спектрограммы S и определение двух ключевых аспектов, а именно, наличие речевой активности (Voice Activity Detection – VAD) и признаков смены диктора (Speaker Change Detection – SCD).

Обозначим сверточный модуль как C_{CNN}, который преобразует входную спектрограмму S в последовательность высокоуровневых признаков $H_{speech} = (h_1, h_2, \dots, h_{T'})$, где каждый компонент последовательности $h_t \in R^D$ является вектором признаков размерности D и извлечён из соответствующего временного интервала. Здесь $T' <= T$ число временных шагов после операций субдискретизации (например, из-за шага свёртки больше 1), которое может быть меньше исходного числа кадров T. Таким образом, последовательность H_{speech} сохраняет временную упорядоченность признаков, но с пониженным разрешением.

$$H_{speech} = C_{CNN}(S), \tag{1}$$

Над этим представлением обучаются два параллельных полносвязных классификатора [16]:

F_{VAD}: $h_t \to \{0, 1\}$, где 1 – наличие речи в сегменте вокруг t.

F_{SCD}: $(h_t, h_{t+1}) \to \{0, 1\}$, где 1 – вероятность смены говорящего между t и $t + 1$.

Выходом этого модуля является разбиение исходного аудиопотока $X(t)$ на сегменты $Seg_1, Seg_2, \dots, Seg_K$, внутри которых речевая активность присутствует и диктор, предположительно, не меняется. Каждый сегмент Seg_k ассоциирован с соответствующим фрагментом признаков $H_k \subseteq H_{speech}$.

Вторым модулем является «Модуль идентификации диктора с механизмом внимания к эталонным голосам». Для каждого сегмента Seg_k необходимо определить, какому из целевых дикторов (сотрудников) он принадлежит. Вместо построения отдельной модели для каждого диктора мы используем парадигму мета-обучения (meta-learning) и механизм внимания [17].

Пусть, имеются эталонные записи голосов N целевых дикторов $D_1, D_2, \dots, D_N$. Для каждого диктора $D_i, i = 1(1)N$ из его эталонной записи (например, несколько произнесенных фраз) извлекаются эталонные эмбэддинги голоса. Для получения эталонных эмбеддингов каждого диктора в предложенной архитектуре используется предобученная модель Wav2Vec2, представляющая собой трансформерную архитектуру, обученную на крупномасштабных корпусах для задач распознавания речи [18]. Данная модель преобразует входной аудиосигнал в последовательность высокоуровневых признаков, эффективно кодирующих акустические характеристики голоса. Процедура извлечения эмбеддинга для каждого эталонного аудиофайла включает несколько последовательных этапов. Исходный сигнал приводится к единой частоте дискретизации 16 кГц и подвергается нормализации для устранения вариаций громкости. Нормализованный сигнал подаётся на вход модели Wav2Vec2, которая возвращает тензор признаков размерности T × G, где T соответствует

количеству временных шагов, а G – размерности скрытого представления (для конфигурации base она составляет 768). Для получения фиксированного векторного представления всего аудиофрагмента выполняется операция усреднения по временной оси (mean pooling), в результате чего формируется вектор размерности G. Заключительным этапом является L2-нормализация полученного вектора, обеспечивающая устойчивость при последующих вычислениях косинусного сходства.

Таким образом, каждый эталонный файл диктора преобразуется в эмбеддинг – точку в пространстве фиксированной размерности, отражающую тембральные характеристики голоса. При наличии нескольких эталонных записей для одного диктора они образуют множество векторов, которое может быть использовано как непосредственно в механизме внимания, так и для вычисления усреднённого прототипа. В предложенной архитектуре механизм внимания оперирует всем множеством эталонных эмбеддингов для каждого диктора, что позволяет учитывать естественную вариативность произношения и повышает робастность идентификации в условиях ограниченных данных.

Пусть $E_i = e_i^1, e_i^2, \dots, e_i^M$– множество M_i векторов-эмбэддингов диктора i, полученных с помощью предобученной сети (например, на основе x-vector архитектуры) [19]. Для сегмента Seg_k с признаками $H_k = (h_1, \dots, h_L)$, где L – длина сегмента во временных шагах после свёртки, в работе используется механизм внимания для вычисления взвешенного контекстного вектора, ориентированного на эталонные голоса.

Для каждого вектора признаков h_l из сегмента Seg_k вычисляется его совместимость со всеми эталонными эмбеддингами e_j^mс помощью функции $score\left(h_l, e_j^m\right)$, где j – номер диктора $(j = 1, \dots, N)$, а m – номер эталонной записи $\left(m = 1, \dots, M_j\right)$. В качестве функции совместимости используется скалярное произведение (при этом векторы предполагаются нормированными). Затем веса внимания $a_{l,j,m}$ получаются применением функции *softmax* по всем дикторам и их эталонам, что гарантирует их нормировку. Таким образом, элементы $a_{l,j,m}$ образуют трёхмерный тензор, отражающий вклад каждого эталонного эмбеддинга каждого диктора e_j^m в описание текущего фрагмента речи h_l:

$$a_{l,j,m} = \frac{exp\left(score\left(h_l, e_j^m\right)\right)}{\sum_{p=1}^{N}\sum_{q=1}^{M_p} exp\left(score\left(h_l, e_p^q\right)\right)} \quad (2)$$

где M_p – количество эталонных эмбеддингов для диктора D_p, p – индекс перебора всех дикторов, а q – индекс эталонной записи для диктора p.

Затем для каждого диктора D_j вычисляется вектор внимания, элементы которого c_k^j, рассчитываются на основе признаков сегмента Seg_k:

$$c_k^j = \sum_{l=1}^{L}\sum_{m=1}^{M_j} a_{l,j,m} * e_j^m \quad (3)$$

где M_j – число эталонных эмбеддингов диктора D_j, j – индекс целевого диктора, а k – индекс сегмента.

Здесь e_j^m эталонные эмбеддинги диктора D_j, а веса $a_{l,j,m}$ определяют вклад каждого эталонного вектора в описание текущего сегмента. Итоговый вектор z_k диктор-зависимых признаков для сегмента Seg_k формируется как конкатенация усредненных признаков сегмента $avg(H_k)$ и векторов внимания векторов внимания c_k^j вычисленных для каждого диктора D_j:

$$z_k = [avg(H_k); c_k^1; c_k^2; \dots; c_k^N] \quad (4)$$

где H_k – усреднённый по времени вектор признаков сегмента Seg_k, а N – общее число целевых дикторов.

Вектор z_k подается на полносвязный классификатор $F_{SPK}: z_k \to R^N$ с N выходами (по числу дикторов), который определяет вероятность принадлежности сегмента k каждому из

целевых дикторов. Идентифицированный диктор D_k для сегмента k определяется как $argmax\left(F_{SPK}(z_k)\right)$.

Третьим модулем является «Модуль транскрибации на основе преобразователя (Transformer)». Для непосредственного преобразования аудио в текст сегмента Seg_kиспользуется архитектура типа Transformer-Encoder [20], которая показала высокую эффективность в ASR. Однако, чтобы улучшить распознавание для конкретных дикторов, мы конкретизируем работу декодера на идентифицированном дикторе. Признаки H_k сегмента Seg_k обрабатываются энкодером преобразователя:

$$O_k = Transformer\_Encoder(H_k) \quad (5)$$

Декодер преобразователя, который генерирует последовательность текстовых токенов $Y_k = (y_1, y_2, \dots, y_T)$, модифицируется таким образом, что его начальное скрытое состояние инициализируется (conditioned) вектором эмбэддинга идентифицированного диктора e_{d_k} (усредненный эталонный эмбэддинг для d_k), где T – это длина сгенерированной последовательности, определяемая моментом появления специального токена конца последовательности (<eos>) или достижением максимальной длины (в работе T ≥ 500). Это позволяет декодеру адаптировать стиль генерации под акустические особенности конкретного говорящего.

$$y_t = Transformer\_Decoder\left(y_{t-1}, O_k, e_{d_k}\right) \quad (6)$$

Обучение сети проводится в два этапа.

Предварительное обучение: Модуль сегментации предобучается на задаче VAD/SCD. Модуль идентификации диктора предобучается на большом наборе данных с множеством дикторов для обучения качественного эмбединг-слоя. Модуль транскрибации предобучается на массовом ASR-корпусе (например, LibriSpeech).

Тонкая настройка (Fine-tuning): Вся архитектура целиком дообучается на целевом наборе данных, содержащем записи совещаний с участием N целевых сотрудников. Используется комбинированная функция потерь $L_{total} = \alpha * L_{ASR} + \beta * L_{SPK}$, где L_{ASR} – распознавания речи (Connectionist Temporal Classification – CTC или Cross-Entropy), L_{SPK} – идентификации диктора (Cross-Entropy).

Результаты и обсуждение. В результате исследования разработана формальная архитектура нейронной сети (рисунок 1 и таблица 1). Основные теоретические результаты заключаются в следующем:

1. Предложена трехмодульная математическая модель обработки многопользовательского аудиопотока, формализованная системой уравнений (1-6);
2. Разработан и формально описан механизм идентификации диктора на основе внимания к эталонным эмбэддингам, который позволяет эффективно использовать обучения с малым количеством примеров;
3. Предложен метод условного управления декодером транскриптора на основе идентифицированного диктора. Как показано в ряде исследований [21–22], conditioning декодера на информацию о дикторе позволяет модели фокусироваться на акустических характеристиках целевого говорящего и улучшает точность распознавания, особенно в условиях перекрывающейся речи и множества дикторов;
4. Формализован двухэтапный процесс обучения, обеспечивающий устойчивость модели при ограниченности целевых данных.

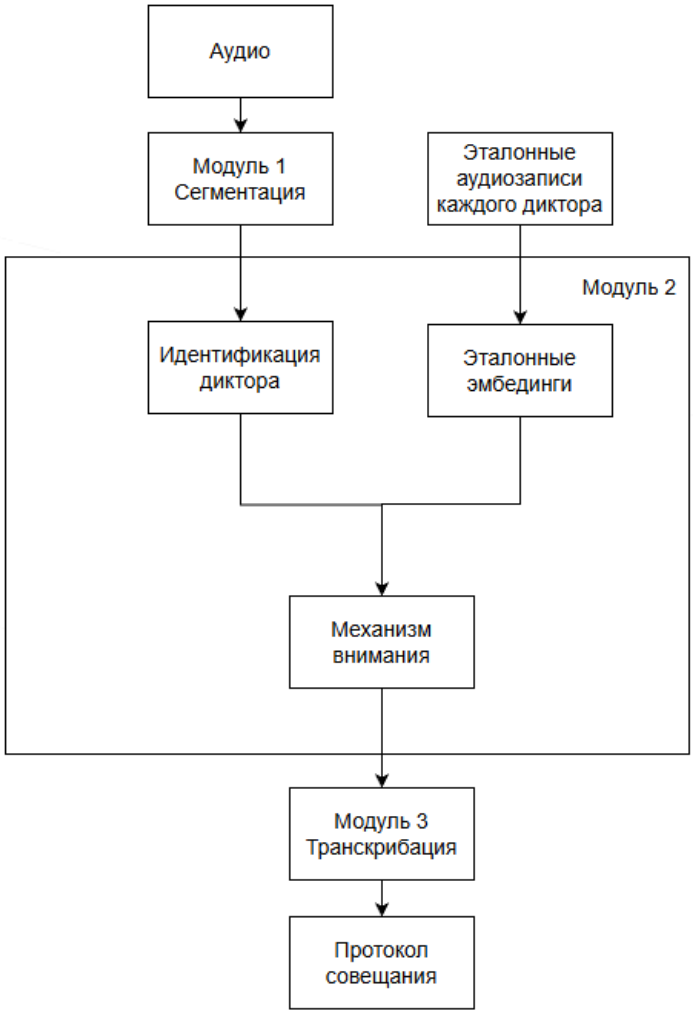

Рис. 1. Общая архитектура предлагаемой гибридной нейронной сети

Таблица 1. Сравнение функций модулей предлагаемой архитектуры

Модуль	Программный аппарат	Входные данные	Выходные данные
1. Сегментации	Глубокая CNN, VAD, SCD	Мел-спектрограмма S	Сегменты речи Seg_k, признаки H_k
2. Идентификации диктора	Механизм внимания, мета-обучение	H_k, эталонные эмбэддинги E_i	Идентификатор диктора d_k
3. Транскрибации	Transformer-Encoder/Decoder	H_k, e_{d_k}	Текстовая транскрипция Y_k

Для оценки эффективности предложенной трёхмодульной гибридной архитектуры были проведены эксперименты на реальных записях корпоративных совещаний общей длительностью более 20 часов. В качестве целевых дикторов выступили пять участников дискуссий, для каждого из которых было подготовлено от 7 до 11 эталонных аудиозаписей длительностью 10–30 секунд. Обучение проводилось в режиме тонкой настройки предобученных компонентов (Wav2Vec2, Whisper) на собранном корпусе.

В таблице 2 представлены полученные метрики качества для каждого модуля разработанной системы.

Таблица 2. Метрики качества модулей гибридной архитектуры

Модуль	Метрика	Значение
Сегментация	VAD Accuracy	0,7940
	VAD F1-score	0,8852
	SCD Precision	0,6245
	SCD Recall	0,4167
	SCD F1-score	0,5000
Идентификация диктора	Speaker Accuracy	0,6254

Модуль сегментации продемонстрировал высокую эффективность в задаче обнаружения речевой активности (VAD F1 = 0,8852), что подтверждает корректность выбора глубокой свёрточной сети для выделения универсальных речевых признаков. Высокий VAD Accuracy (0,7940) свидетельствует о способности модели надёжно выделять речь в условиях реальных совещаний.

В задаче детекции смены диктора (SCD) наблюдается характерный для таких задач дисбаланс классов, при умеренном значении полноты (Recall = 0,4167) точность детекции (Precision = 0,6245) указывает на то, что примерно 62% определяемых смен спикера являются истинными, однако при этом часть реальных смен не идентифицируется. Данный эффект объясняется разреженностью событий смены диктора в обучающей выборке. Улучшение этой метрики требует применения методов борьбы с дисбалансом, таких как взвешенные функции потерь или аугментация данных, что выходит за рамки данной работы.

Модуль идентификации диктора, реализующий предложенный механизм внимания к эталонным эмбеддингам, показал результат 62,5% точности при идентификации пяти дикторов. Данный результат следует признать высоким, учитывая следующие факторы:

- обучение проводилось в условиях few-shot learning (до 11 эталонных записей на диктора);
- предложенная архитектура не требует переобучения при добавлении новых дикторов;
- идентификация выполнялась в условиях естественных диалогов с фоновым шумом и перекрытием речи.

Тем не менее, полученное значение существенно превосходит случайное угадывание (20%) и подтверждает эффективность предложенного подхода с конкатенацией усреднённых признаков сегмента и взвешенных сумм векторов внимания.

Транскрибация в разработанной архитектуре изначально предполагалась, как встроенный модуль на базе трансформера. Однако экспериментально установлено, что для достижения приемлемого качества распознавания речи требуются значительно большие объёмы текстовых транскрипций (50–100 часов), что не всегда доступно в корпоративной среде. В связи с этим в практической реализации системы принято решение использовать гибридный подход: модули сегментации и идентификации работают в предложенной архитектуре, а транскрибация каждого выделенного сегмента выполняется предобученной моделью Whisper. Это обеспечивает высокое качество распознавания речи без необходимости дополнительного обучения и полностью соответствует корпоративным сценариям использования.

Заключение. Полученные результаты дают основание утверждать, что разработанная архитектура работоспособна в обозначенных ранее условиях.

Интеграция задач в единой модели обеспечила согласованную работу модулей сегментации и идентификации. Так, достаточно высокое качество детекции речевой активности (VAD F1 = 0,885) позволяет с приемлемым качеством выделять реплики, а точность идентификации диктора (62,5% для пяти участников) демонстрирует, что механизм внимания к эталонным эмбеддингам успешно обучается и способен сопоставлять признаки

сегмента с голосовыми эталонами, что подтверждает, необходимость совместного обучения, позволяющего избегать накопления ошибок, характерного для раздельных конвейеров.

Устойчивость к ограниченным данным является ключевым преимуществом предложенного подхода, так как применение механизма внимания к эталонным записям является эффективной реализацией few-shot learning, ввиду того, что модель не требует сбора больших размеченных наборов для каждого нового диктора. Всего 7-11 эталонных записей на одного спикера обеспечивают точность идентификации равной 62,5%, что значительно выше случайного уровня (20%). Это подтверждает применимость архитектуры в корпоративной среде, где количество данных для новых сотрудников весьма ограничено.

Детекция смены диктора (SCD) показала умеренные результаты (F1 ≈ 0,5). Причина этого – сильный дисбаланс классов: в реальных диалогах смены происходят относительно редко, и модель испытывает трудности с их обнаружением. Для улучшения этого компонента целесообразно использовать взвешенные функции потерь и методы аугментации (улучшения, расширения или добавления новых качеств данных), например, искусственное создание смен путём склейки сегментов разных дикторов.

Транскрибация речи в исходной архитектуре предполагала встроенный декодер на трансформере, однако проведенные исследования показали, что для качественного обучения потребуется значительное количество размеченных текстовых данных (более 50 часов), что не всегда доступно. В связи с этим в практической реализации было принято решение использовать внешнюю ASR-систему Whisper для распознавания текста выделенных сегментов. Такой гибридный подход сохраняет модульность, не снижая качества идентификации, и позволяет получать точные транскрипции без дополнительного обучения.

Основным ограничением остаётся вычислительная сложность механизма внимания при большом числе дикторов, однако для корпоративных совещаний с типичным числом участников до 10 это не столь критично. Кроме того, качество идентификации напрямую зависит от чистоты эталонных записей – в реальных условиях может потребоваться предварительная очистка аудио.

Список источников

1. Ермоленко Т.В. Система автоматического распознавания слитной русской речи на основе глубоких нейросетей / Т.В. Ермоленко, Я.С. Пикалёв // Речевые технологии, 2021. – № 1-2. – С. 3-18. – DOI: 10.58633/2305-8129_2021_1-2_3.
2. Mujtaba D., Mahapatra N., Arney M. et al. Lost in transcription: identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 4795-4809, DOI: 10.18653/v1/2024.naacl-long.269.
3. Cornell S., Wiesner M.S., Watanabe S. et al. The CHiME-7 DASR challenge: distant meeting transcription with multiple devices in diverse scenarios. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023), 2023, DOI: 10.21437/CHiME.2023-1.
4. Sharma M., Joshi S., Chatterjee T. et al. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. Neurocomputing, 2022, vol. 494, pp. 116-131, DOI: 10.1016/j.neucom.2022.04.084.
5. Суин А.Е. Использование каскадов моделей для резюмирования встреч / А.Е. Суин, В.О. Серебряков, Т.Ю. Чернышева // Промышленный искусственный интеллект (ПИИ'2025): Всероссийская научно-практическая конференция. – СПб.: ПОЛИТЕХ-ПРЕСС, 2025. – С. 439–447. – DOI: 10.18720/SPBPU/2/id25-548.
6. Khadar K.V. A., Kumar R.K. S., Sreekanth N.S. Adapting to noise in forensic speaker verification using GMM-UBM I-vector method in high-noise backgrounds. Computational Sciences and Sustainable Technologies (Communications in Computer and Information Science), 2024, pp. 275-287, DOI: 10.1007/978-3-031-50993-3_22.
7. Abd El-Moneim S., El-Shafai W., Hammam H. et al. Effect of interference on text-independent speaker recognition based on deep learning. Multimedia Tools and Applications, 2025, vol. 84, pp. 47425–47458, DOI: 10.1007/s11042-024-19493-1.

8. An K., Zhang S., Yan Z. Are transformers in pre-trained LM a good ASR encoder? an empirical study. arXiv (Cornell University), 2024, DOI: 10.48550/arXiv.2409.17750.
9. Hirano Y., Sakti S. SC-SOT: conditioning the decoder on diarized speaker information for end-to-end overlapped speech recognition. arXiv (Cornell University), 2025, DOI: 10.48550/arXiv.2506.12672.
10. Jakubec M., Jarina R., Jakubcova E. et al. Deep speaker embeddings for speaker verification: review and experimental comparison. Engineering Applications of Artificial Intelligence, 2023, vol. 123, art. 107025, DOI: 10.1016/j.engappai.2023.107232.
11. Kriman S., Beliaev S., Ginsburg B. et al. Quartznet: deep automatic speech recognition with 1D time-channel separable convolutions. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6124-6128, DOI: 10.1109/ICASSP40776.2020.9053889.
12. Latif S., Zaidi A., Cuayahuitl H. et al. Transformers in speech processing: a survey. arXiv (Cornell University), 2025, DOI: 10.48550/arXiv.2303.11607.
13. Mavaddati S. A voice activity detection algorithm using deep learning in time–frequency domain. Neural Computing and Applications, 2025, vol. 37, pp. 2581–2595, DOI: 10.1007/s00521-024-10795-x.
14. McFee B., Raffel C., Liang D. et al. Librosa: audio and music signal analysis in python. 14th Conference: Python in Science Conference, 2015, pp. 18-25, DOI: 10.25080/Majora-7b98e3ed-003.
15. Polok A., Klemen D., Kocour M. et al. DiCoW: diarization-conditioned whisper for target speaker automatic speech recognition. Computer Speech and Language, 2026, vol. 95, no. 1, pp. 1-19, DOI: 10.1016/j.csl.2025.101841.
16. Rafidison M.A., Rakotomihamina A.H., Rajaonarison F.T.M. et al. Intervention of light convolutional neural network in document survey form processing. Multimedia Tools and Applications, 2023, vol. 82, pp. 32583–32605, DOI: 10.1007/s11042-023-16076-4.
17. Snyder D., Garcia-Romero D., Sell G. et al. X-vectors: robust DNN embeddings for speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, DOI: 10.1109/ICASSP.2018.8461375.
18. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477, 2020, DOI: 10.48550/arXiv.2006.11477.
19. Su H., Zhao D., Dang L. et al. A multitask learning framework for speaker change detection with content information from unsupervised speech decomposition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8087-8091, DOI: 10.1109/ICASSP43922.2022.9746116.
20. Szwoch G. Crosstalk suppression in a multi-channel, multi-speaker system using acoustic vector sensors. MDPI Sensors, 2025, vol. 23, issue 21, pp. 6731, DOI: 10.3390/s25216731.
21. Wang W., Zhao H., Yang Y. et al. Few-shot short utterance speaker verification using meta-learning. PeerJ Computer Science, 2023, 9:e1276, DOI: 10.7717/peerj-cs.1276.
22. Cui C., Sheikh I., Sadeghi M., Vincent E. End-to-end multichannel speaker-attributed ASR: speaker guided decoder and input feature analysis. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2023, DOI: 10.1109/ASRU57964.2023.10389729.

Потапов Арсений Денисович. *Студент магистратуры Иркутского национального исследовательского технического университета, эксперт по цифровым технологиям аудита Байкальского банка ПАО Сбербанк. ORCID: 0009-0008-5949-1293, arseny.potapov2012@yandex.ru, 664074, г. Иркутск, ул. Лермонтова 83.*

Лукьянов Никита Дмитриевич. *Кандидат технических наук, руководитель центра электронного обучения, доцент лаборатории сетевых систем и ИТ-инфраструктуры, Иркутского национального исследовательского технического университета. Направления исследований: алгоритмы машинного обучения и искусственного интеллекта, автоматизация в промышленности, математическое моделирование. AuthorID: 732737, SPIN: 8971-7440, ORCID: 0000-0002-8826-3424, lukyanovnd@ex.istu.edu, 664074, г. Иркутск, ул. Лермонтова 83.*

UDC 004.934.2
DOI:10.25729/ESI.2026.43.3.003

Adaptive hybrid neural network architecture with attention for speaker identification in corporate meetings

Arseniy D. Potapov[1], Nikita D. Lukyanov[2]

[1]PJSC Sberbank of Russia, Russia, Irkutsk

[2]Irkutsk National Research Technical University, Russia, Irkutsk, *lukyanovnd@istu.edu*

Abstract. The relevance of this research stems from the growing need to automate corporate meeting minutes, a labor-intensive and time-consuming process with a high risk of errors. Current speech recognition systems (ASR – Automatic Speech Recognition) perform poorly in multi-user conversations, with overlapping voices, background noise, and the use of specialized vocabulary. Speaker identification (SID – Speaker Identification) is particularly challenging when there are few reference recordings per employee, which is typical in an office environment. Acoustic environments also pose additional challenges: echoing offices, simultaneous speech, and poor voice separation. Therefore, the goal of this study is to develop a hybrid neural network model for accurate transcription and reliable speech attribution in group discussions. This paper analyzes the weaknesses of ASR systems and SID methods and proposes a three-tier architecture: audio stream segmentation (VAD – Voice Activity Detection, SCD – Speaker Change Detection), voice embedding identification with a reference attention mechanism, and speaker-aware transcription. A mathematical justification for the key components is also provided. The theoretical significance lies in the development of an attention-based SID method that effectively solves few-shot learning problems. Unlike traditional approaches that require large labeled corpora, the new architecture learns to match speech fragments with a small number of reference recordings. The effectiveness of the method has been empirically confirmed using real corporate meeting recordings (over 20 hours, five participants). The segmentation module demonstrated a VAD F1 score of 0.885, and the speaker identification module achieved an accuracy of 62.5%, which is three times higher than chance (20%). Transcription was performed using the external Whisper system, which guarantees high recognition accuracy. The results enable the development of a system for automatic meeting minutes, and the architecture supports a large number of participants and is easily adapted to various communication formats: from negotiations to client meetings. The functionality of the hybrid neural network model can be used to create various tools for corporate meeting minutes.

Keywords: Automatic speech recognition, speaker identification, hybrid neural network architecture, attention mechanism, voice embeddings, audio stream segmentation, transcription, corporate meetings

References

1. Ermolenko T.V., Pikalev Ya.S. Sistema avtomaticheskogo raspoznavaniya slitnoy russkoy rechi na osnove glubokikh neyrosetey [System for automatic continuous Russian speech recognition based on deep neural networks]. Rechevyye tekhnologii [Speech Technologies], 2021, no. 1-2, pp. 3-18, DOI: 10.58633/2305-8129_2021_1-2_3.
2. Mujtaba D., Mahapatra N., Arney M. et al. Lost in transcription: identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 4795-4809, DOI: 10.18653/v1/2024.naacl-long.269.
3. Cornell S., Wiesner M.S., Watanabe S. et al. The CHiME-7 DASR challenge: distant meeting transcription with multiple devices in diverse scenarios. 7th International Workshop on Speech Processing in Everyday Environments (CHiME 2023), 2023, DOI: 10.21437/CHiME.2023-1.
4. Sharma M., Joshi S., Chatterjee T. et al. A comprehensive empirical review of modern voice activity detection approaches for movies and TV shows. Neurocomputing, 2022, vol. 494, pp. 116-131, DOI: 10.1016/j.neucom.2022.04.084.
5. Suin A.E., Serebryakov V.O., Chernysheva T.Yu. Ispol'zovanie kaskadov modeley dlya rezyumirovaniya vstrech [Using cascades of models for meeting summarization]. Promyshlennyy iskusstvennyy intellekt (PII'2025): Vserossiyskaya nauchno-prakticheskaya konferentsiya [Industrial Artificial Intelligence (IAI'2025): All-Russian Scientific and Practical Conference], St. Petersburg, POLYTECH-PRESS, 2025, pp. 439-447. DOI: 10.18720/SPBPU/2/id25-548.
6. Khadar K.V. A., Kumar R.K. S., Sreekanth N.S. Adapting to noise in forensic speaker verification using GMM-UBM I-vector method in high-noise backgrounds. Computational Sciences and Sustainable Technologies (Communications in Computer and Information Science), 2024, pp. 275-287, DOI: 10.1007/978-3-031-50993-3_22.

7. Abd El-Moneim S., El-Shafai W., Hammam H. et al. Effect of interference on text-independent speaker recognition based on deep learning. Multimedia Tools and Applications, 2025, vol. 84, pp. 47425–47458, DOI: 10.1007/s11042-024-19493-1.
8. An K., Zhang S., Yan Z. Are transformers in pre-trained LM a good ASR encoder? an empirical study. arXiv (Cornell University), 2024, DOI: 10.48550/arXiv.2409.17750.
9. Hirano Y., Sakti S. SC-SOT: conditioning the decoder on diarized speaker information for end-to-end overlapped speech recognition. arXiv (Cornell University), 2025, DOI: 10.48550/arXiv.2506.12672.
10. Jakubec M., Jarina R., Jakubcova E. et al. Deep speaker embeddings for speaker verification: review and experimental comparison. Engineering Applications of Artificial Intelligence, 2023, vol. 123, art. 107025, DOI: 10.1016/j.engappai.2023.107232.
11. Kriman S., Beliaev S., Ginsburg B. et al. Quartznet: deep automatic speech recognition with 1D time-channel separable convolutions. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6124-6128, DOI: 10.1109/ICASSP40776.2020.9053889.
12. Latif S., Zaidi A., Cuayahuitl H. et al. Transformers in speech processing: a survey. arXiv (Cornell University), 2025, DOI: 10.48550/arXiv.2303.11607.
13. Mavaddati S. A voice activity detection algorithm using deep learning in time–frequency domain. Neural Computing and Applications, 2025, vol. 37, pp. 2581–2595, DOI: 10.1007/s00521-024-10795-x.
14. McFee B., Raffel C., Liang D. et al. Librosa: audio and music signal analysis in python. 14th Conference: Python in Science Conference, 2015, pp. 18-25, DOI: 10.25080/Majora-7b98e3ed-003.
15. Polok A., Klemen D., Kocour M. et al. DiCoW: diarization-conditioned whisper for target speaker automatic speech recognition. Computer Speech and Language, 2026, vol. 95, no. 1, pp. 1-19, DOI: 10.1016/j.csl.2025.101841.
16. Rafidison M.A., Rakotomihamina A.H., Rajaonarison F.T.M. et al. Intervention of light convolutional neural network in document survey form processing. Multimedia Tools and Applications, 2023, vol. 82, pp. 32583–32605, DOI: 10.1007/s11042-023-16076-4.
17. Snyder D., Garcia-Romero D., Sell G. et al. X-vectors: robust DNN embeddings for speaker recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, DOI: 10.1109/ICASSP.2018.8461375.
18. Baevski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477, 2020, DOI: 10.48550/arXiv.2006.11477.
19. Su H., Zhao D., Dang L. et al. A multitask learning framework for speaker change detection with content information from unsupervised speech decomposition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8087-8091, DOI: 10.1109/ICASSP43922.2022.9746116.
20. Szwoch G. Crosstalk suppression in a multi-channel, multi-speaker system using acoustic vector sensors. MDPI Sensors, 2025, vol. 23, issue 21, pp. 6731, DOI: 10.3390/s25216731.
21. Wang W., Zhao H., Yang Y. et al. Few-shot short utterance speaker verification using meta-learning. PeerJ Computer Science, 2023, 9:e1276, DOI: 10.7717/peerj-cs.1276.
22. Cui C., Sheikh I., Sadeghi M., Vincent E. End-to-end multichannel speaker-attributed ASR: speaker guided decoder and input feature analysis. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2023, DOI: 10.1109/ASRU57964.2023.10389729.

Potapov Arseniy Denisovich. *Master's student at Irkutsk National Research Technical University, expert in digital technologies of audit of Baikal Bank of Sberbank. ORCID: 0009-0008-5949-1293, arseny.potapov2012@yandex.ru, 664074, Irkutsk, 83 Lermontov Street.*

Lukyanov Nikita Dmitriyevich. *PhD in Engineering, Head of the E-Learning Center, Associate Professor at the Laboratory of Network Systems and IT Infrastructure at Irkutsk National Research Technical University. Research areas: machine learning and artificial intelligence algorithms, industrial automation, mathematical modeling. AuthorID: 732737, SPIN: 8971-7440, ORCID: 0000-0002-8826-3424, lukyanovnd@ex.istu.edu, 664074, Irkutsk, 83 Lermontov Street.*

Статья поступила в редакцию 07.04.2026; одобрена после рецензирования 25.05.2026; принята к публикации 29.07.2026.

The article was submitted 04/07/2026; approved after reviewing 05/25/2026; accepted for publication 07/29/2026.