

ПСИХОЛОГИЯ ТРУДА И ОРГАНИЗАЦИОННАЯ ПСИХОЛОГИЯ

Статья поступила в редакцию: 19.09.2026

Статья принята к публикации: 24.09.2026

Статья опубликована: 08.10.2026

УДК 159.98:004.8

DOI: 10.32743/UniPsy.2026.148.10.23454

Программный комплекс анализа экстралингвистических характеристик речи в психологическом консультировании: архитектура и принцип трёхуровневой интерпретируемости

Шанин Александр Владимирович

докторант

институт Смарт

Российская Федерация, г. Москва

E-mail: Otis13@ya.ru

A software system for analysing extralinguistic speech characteristics in psychological counselling: architecture and the principle of three-level interpretability

Shanin Aleksandr Vladimirovich

Smart Institute

Russia, Moscow

Аннотация

Распространение больших языковых моделей в психологической практике обострило проблему доверия к автоматически сформированным выводам: связный и профессионально звучащий текст не содержит указания на то, из каких данных он получен, а ответственность за сказанное клиенту остаётся за специалистом. В статье описан программный комплекс «Оракул Лайт», предназначенный для анализа экстралингвистических характеристик речи в индивидуальном психологическом консультировании, и обоснован положенный в его основу принцип трёхуровневой интерпретируемости. Консультативная сессия рассматривается как объект, имеющий акустический, структурный и аналитический уровни. Интерпретируемость понимается не как объяснение внутреннего устройства модели, а как свойство выходного текста: возможность проследить любой вывод от формулировки гипотезы через её адрес во фрагменте сессии до числового показателя, вычисленного детерминированным алгоритмом. Показано, как этот принцип закреплён в архитектуре комплекса. Языковая модель не

Библиографическое описание: Шанин А.В. Программный комплекс анализа экстралингвистических характеристик речи в психологическом консультировании: архитектура и принцип трёхуровневой интерпретируемости // Universum: психология и образование : электрон. научн. журн. 2026. 10(148). URL: <https://7universum.com/ru/psy/archive/item/23454>

получает аудиозапись и не выполняет измерений; интерпретационная версия принимается только при двойной опоре на реплику клиента и акустический показатель того же фрагмента; каждая версия сопровождается явным уровнем уверенности и условием, которое её опровергает; противоречащие сигналы подлежат обязательному называнию. Рассмотрено практическое значение подхода для специалиста в рамках концепции «ИИ как ассистент профессиональной рефлексии», а также ограничения, связанные с пилотным статусом апробации, тематической однородностью выборки и зависимостью результатов от качества автоматического разделения говорящих.

Abstract

The spread of large language models in psychological practice has sharpened the problem of trust in automatically generated conclusions: a coherent, professionally worded text gives no indication of the data it was derived from, while responsibility for what is said to the client remains with the practitioner. The article describes Oracul Lite, a software system for analysing extralinguistic speech characteristics in individual psychological counselling, and substantiates the principle of three-level interpretability underlying it. A counselling session is treated as an object with acoustic, structural and analytical levels. Interpretability is understood not as an explanation of the model's internal workings but as a property of the output text: the possibility of tracing any conclusion from the wording of a hypothesis, through its location in a session fragment, to a numerical indicator computed by a deterministic algorithm. The article shows how this principle is embedded in the architecture of the system. The language model receives no audio recording and performs no measurements; an interpretative version is accepted only with dual support, a client utterance and an acoustic indicator from the same fragment; every version carries an explicit confidence level and a condition that would refute it; contradictory signals must be reported. The practical value of the approach is discussed within the concept of AI as an assistant to professional reflection, together with limitations related to the pilot status of the evaluation, the thematic homogeneity of the sample and the dependence of the results on the quality of automatic speaker diarization.

Ключевые слова: экстралингвистические характеристики речи, психологическое консультирование, большие языковые модели, трёхуровневая интерпретируемость, экстралингвистический профиль консультативной сессии, верифицируемость выводов, профессиональная рефлексия.

Keywords: extralinguistic speech characteristics, psychological counselling, large language models, three-level interpretability, extralinguistic profile of a counselling session, verifiability of conclusions, professional reflection.

Введение

Получив от программы текст о собственном клиенте, опытный консультант спрашивает прежде всего не о том, насколько текст глубок, а о том, откуда он взялся. Вопрос этот не праздный: всё, что специалист затем скажет клиенту или положит в план следующей встречи, ложится на его профессиональную ответственность, и ссылка на «так посчитала система» её не снимает.

С появлением больших языковых моделей вопрос стал заметно острее. Модели пишут гладко, уверенно и в узнаваемом профессиональном регистре – и как раз поэтому оценить написанное труднее, чем прежде. Применение таких моделей в психологии стало предметом отдельного обсуждения [4]. Лоуренс и др. [8] в обзоре возможностей и рисков языковых моделей в сфере психического здоровья предупреждают, что неточная информация может оказаться вреднее, чем её отсутствие, особенно когда неточность трудно заметить, и настаивают, что решение о принятии предложений модели остаётся за специалистом. Для консультанта это означает простую вещь: убедительность вывода ничего не говорит о его точности.

Традиция автоматического анализа голоса, существовавшая задолго до языковых моделей, эту проблему не решает, а по-своему воспроизводит. Обзор Эль-Айяди и др. [5] описывает типичную схему распознавания эмоций по речи: из сигнала извлекаются акустические признаки, классификатор соотносит их с одной из заранее заданных категорий. Точность такого распознавания ограничена: по данным Шерера [10], на записях актёрских портретов эмоций она составляет от 55 до 65 %. Хазарика и др. [6] ввели в модель контекст диалога – предшествующие реплики обоих собеседников в диаде, – однако итогом анализа по-прежнему остаётся метка эмоционального класса. В работе Трана и др. [11] по тексту и звуку реальных сессий мотивационного интервьюирования автоматически распознаётся уровень эмпатии консультанта, и результатом снова служит итоговая оценка.

Практику такая метка мало что даёт: с ней можно согласиться или не согласиться, но её нельзя проверить – в ней не указано, на каком отрезке беседы и по какому признаку она получена. Мы исходим из того, что вопрос «как сделать систему точной» применительно к интерпретации не имеет окончательного ответа: два квалифицированных специалиста могут по-разному прочесть один и тот же случай, и оба прочтения будут профессионально обоснованы. Решаемым мы считаем другой вопрос – как сделать выводы системы проверяемыми.

Цель исследования

Цель работы – обосновать принцип трёхуровневой интерпретируемости как способ обеспечить проверяемость выводов системы автоматизированного анализа речи в психологическом консультировании и показать, как этот принцип реализован в архитектуре программного комплекса «Оракул Лайт» (Oracul Lite, Python 3.12 / FastAPI / WhisperX / ruannote.audio, 2025–2026). Техническое устройство комплекса рассматривается здесь не само по себе, а в той мере, в какой оно этот принцип воплощает.

Материал и методы исследования

Материалом служат аудиозаписи индивидуальных консультаций, проводимых автором в рамках собственной практики. Каждая встреча длится от 20 до 30 минут, проходит в диаде «консультант – клиент» и строится по фиксированному скрипту из десяти вопросов, направленных на прояснение запроса. Единый формат сессий – условие сопоставимости акустических данных между записями.

Методологически работа опирается на представление о консультативной сессии как о трёхуровневом объекте. Первый уровень, акустический, включает объективно измеримые характеристики: темп речи, вариативность основного тона, громкость, длительность высказываний и долю пауз. Здесь ничего не интерпретируется, здесь измеряют. Второй уровень, структурный, отвечает на вопрос, кто говорит и в какой части беседы: он разделяет речь психолога и клиента и членит сессию на смысловые блоки. Третий уровень, аналитический, строит из измерений и структуры профессиональные гипотезы. Совокупность акустических показателей, разнесённых по структуре беседы, мы называем экстралингвистическим профилем консультативной сессии; принципиально, что этот профиль описывает процесс, а не состояние человека.

Обработка записи устроена как конвейер, где за каждым этапом закреплена ровно одна операция. Речь переводится в текст средствами WhisperX – надстройки над моделью распознавания речи Whisper, обученной на большом массиве слабо размеченных данных [9]. Исходная модель не выдаёт пословных временных меток, и WhisperX получает их принудительным выравниванием фонемного распознавания с аудиосигналом [1]. Разделение говорящих выполняет пакет ruannote.audio – набор нейросетевых компонентов, включающий определение речевой активности и векторные представления голоса [2]. Роли психолога и клиента назначаются по взвешенной сумме четырёх признаков речевого поведения: доли времени речи, доли слов, средней длины реплики и частоты вопросов. Просодические характеристики извлекаются средствами Praat через программный интерфейс parselmouth [7], энергетические – средствами библиотеки анализа аудио librosa.

Границы смысловых блоков система не угадывает: консультант после каждого вопроса нажимает элемент интерфейса, который фиксирует момент перехода. Адрес любого фрагмента – «ответ на седьмой вопрос» – известен системе точно, а не восстановлен задним числом по тексту.

Отдельного обоснования требует выбор меры вариативности тона. Интуитивно напрашивается простой размах: расстояние от самого низкого значения частоты во фрагменте до самого высокого. Однако алгоритм отслеживания тона время от времени ошибается на октаву, и единственная ошибочная точка раздувает размах до бессмысленных величин: на одной из ранних записей размах составил 318 Гц при медиане около 100 Гц. В действующей версии основной мерой служит стандартное отклонение частоты основного тона, устойчивое к единичным выбросам, а диапазон поиска частоты задаётся с учётом пола говорящего.

Языковая модель подключается только на третьем уровне. Аудиозапись ей не передаётся: на вход подаётся структурированный пакет – транскрипт, разбитый по блокам и размеченный по ролям, агрегированные акустические показатели каждого блока и сводные характеристики сессии. Эта же граница разделяет два вычислительных контура системы. Все измерения выполняются на локальном сервере, откуда аудиозапись и голос как биометрическая характеристика не выходят; внешней языковой модели через промежуточный шлюз передаются только текст и числа. Такое устройство реализует подход «конфиденциальность по архитектуре», при котором защита данных встроена в саму конструкцию системы, а не добавлена к ней после [3].

Результаты исследования и их обсуждение

Термин «интерпретируемость» в литературе по машинному обучению обычно относится к объяснению внутренней работы модели. Мы употребляем его в ином значении: интерпретируемость для нас – свойство не модели, а выходного текста. Внутренности языковой модели остаются закрытыми, и раскрывать их не требуется. Вместо этого к тексту предъявляется требование разложимости: любое утверждение отчёта должно допускать обратный ход – от формулировки гипотезы (аналитический уровень) к конкретному фрагменту сессии, где она основана (структурный уровень), и далее к числу, вычисленному для этого фрагмента (акустический уровень). Три уровня модели сессии превращаются в три ступени проверки, отсюда и название принципа.

У принципа есть простой рабочий критерий. Специалист, располагающий несколькими минутами, берёт из отчёта любое утверждение и пытается найти в транскрипте реплику, на которую оно опирается, и показатель, который его подкрепляет. Если это удаётся, вывод можно принимать или отклонять уже по существу. Если обратный ход обрывается, утверждение не обосновано, как бы убедительно оно ни звучало. Отсюда вытекает требование, которое легко упустить: проследить вывод до числа имеет смысл лишь тогда, когда само число надёжно. Обратный ход к ошибочному измерению создаёт видимость обоснованности там, где её нет, и это хуже прямого отсутствия опоры. Поэтому трёхуровневая интерпретируемость начинается не с языковой модели, а с выбора метрик на акустическом уровне.

Разделение труда между алгоритмами и моделью сужает саму природу возможных ошибок. Ошибиться в измерении модель не может, потому что не измеряет; ошибиться она может только в прочтении. Ошибка прочтения обнаружима, поскольку исходный пакет данных и необработанный ответ модели сохраняются вместе с итоговым текстом отчёта. Инструкция для модели зафиксирована в единой редакции и не менялась на протяжении апробации, иначе сопоставлять отчёты разных сессий было бы некорректно.

Сами по себе эти меры не гарантируют, что текст отчёта будет опираться на данные: соблюсти любую форму можно и при содержательно произвольном наполнении. Поэтому основную нагрузку несёт система требований, предъявляемых к каждому выводу модели. Центральное из них – требование двойной опоры. Версия о причинах, скрытых за запросом клиента, принимается лишь при наличии двух оснований одновременно: дословной или близкой к дословной реплике клиента и акустического показателя, выделяющегося на том же фрагменте сессии. Если голос в этом месте ничего не добавляет к словам, версия допускается с одной, речевой, опорой, но уровень уверенности в ней не может быть выше среднего. Подставлять цифру ради солидности запрещено. Двойная опора переводит интерпретируемость из декларации в условие порождения текста: обратный ход закладывается в вывод в момент его формулировки, а не восстанавливается специалистом потом.

С двойной опорой связан запрет на очевидное. Акустический показатель считается аргументом только тогда, когда сообщает то, чего нельзя было узнать из слов. Формула «много пауз – значит, тема трудная» лишь подписывает цифру и ничего не открывает. Ценно другое – рассогласование между сказанным и тем, как это прозвучало. Приведём условный

пример того уровня, на который ориентирована система. Клиент называет главной своей трудностью отсутствие цели и говорит об этом относительно ровно; зато на вопросе о том, по каким признакам он поймёт, что цель достигнута, речь дробится сильнее всего за сессию, а вариативность тона падает до минимума. Словами клиент этот момент не выделил и прошёл его вскользь. Отсюда рабочая гипотеза: застревание, возможно, связано не с выбором цели, а с отсутствием внутреннего критерия её достижения. Именно такие места консультант, удерживающий в фокусе содержание беседы, рискует пропустить. Показатель, который лишь повторяет вывод, уже очевидный из слов, опасен другим: он создаёт иллюзию проверенности и ослабляет критичность читателя.

Каждая версия сопровождается двумя обязательными элементами – явным уровнем уверенности и указанием на то, что могло бы её опровергнуть. Второй элемент не менее важен, чем первый. Гипотеза с условием опровержения перестаёт быть суждением, которое остаётся только принять или отвергнуть, и становится предметом проверки на следующей встрече: в отчёте ей соответствует конкретный вопрос с пометкой, какую именно версию он проверяет. Проверка тем самым выносится за пределы текста – в живую работу с клиентом, где у специалиста есть собственные средства различения.

Отдельное правило касается формы, в которой числа предъявляются читателю. Консультант не обязан помнить, что означают восемь герц стандартного отклонения тона, поэтому любое значение приводится со сравнением внутри той же сессии: не «вариативность тона 8,4 Гц», а «самый ровный тон за встречу – 8,4 Гц против 15–22 Гц на других темах». Внутрисессионное сравнение решает сразу две задачи: делает цифру понятной без справочника и избавляет от опоры на внешние нормативы, которых для подобных метрик в консультативной речи не существует. Наконец, инструкция запрещает замалчивать противоречащие сигналы: если громкость на фрагменте растёт, а вариативность тона падает, модель обязана назвать оба факта и предложить два прочтения. Выбор удобного сигнала при умолчании о неудобном – самый естественный способ получить гладкий и убедительный текст, поэтому он закрыт прямым запретом.

Структура отчёта подчинена тому же принципу, но выстроена от выводов к данным. Консультант присутствовал на сессии, и пересказ беседы ему не нужен. Отчёт открывается сопоставлением исходной и итоговой формулировок запроса, продолжается конкурирующими версиями того, что может за ним стоять, адресными направлениями работы и опорами для следующей встречи и завершается разделом опорных показателей. Этот последний раздел не вводит новых трактовок: в нём собраны именно те числа, на которые опирались версии, и при каждом помечено, какую версию оно подкрепляет. Круг замыкается внутри самого документа, а в печатной версии отчёта сводная таблица показателей приложена к аналитическому тексту.

Что получает специалист от системы, устроенной подобным образом? Прежде всего не новые сведения о клиенте, а возможность вернуться к прошедшей встрече с другой стороны. Внимание консультанта во время сессии поглощено содержанием разговора, удержанием контакта и выбором следующего шага; форма речи – где клиент замедлился, где замолчал дольше обычного, где голос перестал совпадать со словами – фиксируется в лучшем случае фрагментарно. Отчёт возвращает эту сторону беседы в поле рассмотрения. Мы обозначаем такую роль системы как «ИИ как ассистент профессиональной рефлексии»: система не

сообщает психологу, что происходило на встрече, – он там был, – а задаёт вопрос, близкий по духу к супервизорскому: что происходило вот в этом месте и почему оно звучит иначе, чем соседние?

Трёхуровневая интерпретируемость защищает специалиста от двух противоположных ошибок. Первая – некритичное принятие вывода, звучащего профессионально и уверенно. Вторая, о которой говорят реже, – огульное отвержение: если вывод нельзя проверить, осторожный практик разумно отказывается от него целиком, вместе с тем полезным, что в нём было. Возможность обратного хода позволяет обойтись без обеих крайностей: консультант может принять одну версию, отклонить другую, третью отложить до следующей встречи – и в каждом случае знать, на каком основании он это делает.

Разграничение ответственности формулируется в системе явно. Система отвечает за корректность измерений и за честную маркировку того, на чём держится каждая гипотеза и насколько она уверена. Специалист отвечает за интерпретацию, за выбор того, что из отчёта взять в работу, и за всё, что будет сказано клиенту. Сам клиент отчёт не читает и знакомится с выводами системы в устном обсуждении со специалистом. Инструкция запрещает модели описывать черты личности: одна встреча не даёт для этого оснований, и подкрепить такие суждения системе было бы нечем. Рабочий материал отчёта – темы, противоречия и динамика в пределах конкретной сессии.

Предлагаемый подход отличается и от направления, в котором развивается автоматический анализ записей консультаций. В работах, подобных исследованиям Хазарики и др. [6] и Трана и др. [11], задача ставится как распознавание: определить эмоцию собеседника или уровень эмпатии консультанта. Мы сознательно отказываемся от итоговой оценки в пользу набора проверяемых гипотез с адресами. Точность такого отчёта нельзя выразить одной цифрой, зато каждое его утверждение открыто для профессиональной проверки.

Ограничения подхода назовём прямо. Апробация системы имеет пилотный характер и ведётся на тематически однородной выборке: клиенты обращались с запросами, связанными со смыслом, самореализацией и целеполаганием, и переносить результаты на другие категории запросов пока нельзя. Сессии проводит, а отчёты оценивает один специалист, который одновременно является автором системы, поэтому его оценки не слепые и не свободны от эффекта подтверждения. Самое чувствительное техническое ограничение – зависимость всей цепочки от качества разделения говорящих: если алгоритм отнёс фрагмент речи клиента консультанту, обратный ход сработает безупречно, но приведёт к реплике и показателю не того человека. Трёхуровневая интерпретируемость обеспечивает проверяемость вывода и не защищает от ошибок, допущенных на ранних этапах обработки. Требования к выводам модели заданы инструкцией, а не жёстким программным ограничением, и модель может их нарушить; именно поэтому числа, на которые опирается отчёт, всегда остаются в распоряжении специалиста. Наконец, совпадение выводов системы с наблюдением специалиста само по себе ещё не доказывает, что акустический канал добавляет что-то к анализу одного лишь текста беседы: этот вопрос проверяется отдельной процедурой сопоставления отчётов, построенных на полном и урезанном пакете данных, и выходит за рамки настоящей статьи.

Заключение

Проблему доверия к выводам систем на основе больших языковых моделей в психологической практике мы предлагаем решать не повышением убедительности этих выводов, а обеспечением их проверяемости. Принцип трёхуровневой интерпретируемости понимает интерпретируемость как свойство выходного текста: любое утверждение отчёта должно допускать обратный ход от гипотезы к фрагменту сессии и далее к числу, полученному детерминированным алгоритмом. В программном комплексе «Оракул Лайт» этот принцип закреплён разделением измерения и интерпретации, выбором устойчивых метрик, требованием двойной опоры, явными уровнями уверенности с условиями опровержения и запретом на замалчивание противоречащих сигналов.

Такое устройство оставляет интерпретацию за специалистом, которому она принадлежит и по профессиональным, и по этическим основаниям, и одновременно даёт ему материал для профессиональной рефлексии, недоступный в ходе живой беседы. Дальнейшая работа связана с эмпирической проверкой того, насколько систематически выводы системы согласуются с независимыми критериями, и с оценкой собственного вклада акустического канала.

Список литературы

1. Bain M., Huh J., Han T., Zisserman A. WhisperX: time-accurate speech transcription of long-form audio // Proceedings of Interspeech. – 2023. – P. 4489–4493. – DOI: 10.21437/Interspeech.2023–78.
2. Bredin H., Yin R., Coria J.M., Gelly G., Korshunov P., Lavechin M., Fustes D., Titeux H., Bouaziz W., Gill M.-P. pyannote.audio: neural building blocks for speaker diarization // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP. – 2020. – P. 7124–7128. – DOI: 10.1109/ICASSP40776.2020.9052974.
3. Cavoukian A. Privacy by design: the 7 foundational principles. Toronto: Information and Privacy Commissioner of Ontario [Электронный ресурс]. – 2011. – 5 с. URL: https://student.cs.uwaterloo.ca/~cs492/papers/7foundationalprinciples_longer.pdf (дата обращения: 18.09.2026).
4. Demszky D., Yang D., Yeager D.S., Bryan C.J., Clapper M., Chandhok S., Eichstaedt J.C., Hecht C., Jamieson J., Johnson M., Jones M., Krettek-Cobb D., Lai L., JonesMitchell N., Ong D.C., Dweck C.S., Gross J.J., Pennebaker J.W. Using large language models in psychology. Nature Reviews Psychology. 2023. vol. 2. P. 688–701. DOI: 10.1038/s44159-023-00241-5.
5. El Ayadi M., Kamel M.S., Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition. 2011. vol. 44. no. 3. P. 572–587. DOI: 10.1016/j.patcog.2010.09.020.
6. Hazarika D., Poria S., Zadeh A., Cambria E., Morency L.-P., Zimmermann R. Conversational memory network for emotion recognition in dyadic dialogue videos // Proceedings of the. – 2018. – P. 2122–2132. – DOI: 10.18653/v1/N18-1193.
7. Jadoul Y., Thompson B. de Boer B. Introducing Parselmouth: a Python interface to Praat // Journal of Phonetics. – 2018. – P. 1–15. – DOI: 10.1016/j.wocn.2018.07.001.

8. Lawrence H.R., Schneider R.A., Rubin S.B., Matarić M.J., McDuff D.J., Jones Bell M. The opportunities and risks of large language models in mental health. *JMIR Mental Health*. 2024. vol. 11. P. e59479. DOI: 10.2196/59479.
9. Radford A., Kim J.W., Xu T., Brockman G. McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // *Proceedings of the 40th International Conference on Machine Learning (ICML. – 2023. – P. 28492–28518.*
10. Scherer K.R. Vocal communication of emotion: a review of research paradigms. *Speech Communication*. 2003. vol. 40. no. 1–2. P. 227–256. DOI: 10.1016/S0167-6393(02)00084-5.
11. Tran T., Yin Y., Tavabi L., Delacruz J., Borsari B., Woolley J.D., Scherer S., Soleymani M. Multimodal analysis and assessment of therapist empathy in motivational interviews // *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Paris. – 2023. – P. 406–415. – DOI: 10.1145/3577190.3614105.