Проблемы искусственного интеллекта. 2026. N 3 (42). С. 161-174
Problems of Artificial Intelligence. 2026;3(42):161-174
Системный анализ, управление и обработка информации, статистика
Научная статья

УДК 004.5:004.934 **DOI** 10.24412/2413-7383-2026-3-42-161-174

М. Д. Сидорика[1], А. В. Ниценко[2]
Федеральное государственное бюджетное научное учреждение
«Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артема, 118 б
[1]sidorika.maks@gmail.com. ORCID: 0009-0008-9388-324X.
[2]nav_box@mail.ru. ORCID: 0009-0008-3411-7107

ГИБРИДНАЯ АРХИТЕКТУРА АВТОНОМНОГО ГОЛОСОВОГО АССИСТЕНТА ДЛЯ ЛОКАЛЬНОГО УПРАВЛЕНИЯ ПК

M. D. Sidorika[1], A. V. Nitsenko[2]
Federal State Budgetary Scientific Institution «Institute of Artificial Intelligence Problems»
283048, Donetsk, Artema str, 118-b
[1]sidorika.maks@gmail.com. ORCID: 0009-0008-9388-324X.
[2]nav_box@mail.ru. ORCID: 0009-0008-3411-7107

HYBRID ARCHITECTURE OF AN AUTONOMOUS VOICE ASSISTANT FOR LOCAL PC CONTROL

Актуальность исследования обусловлена потребностью в локальном голосовом управлении ПК без передачи речевых данных во внешние службы. Цель работы – разработать и оценить гибридную архитектуру русскоязычного голосового ассистента. Система объединяет автоматическое распознавание речи на базе Vosk с грамматическим ограничением, модель понимания естественного языка для интерпретации команд, модуль правил и слой контролируемого исполнения. Эксперимент на 1053 аудиопримерах показал рост совместной точности интента и параметров с 83,67 до 96,58 %, снижение доли пословных ошибок с 20,05 до 12,56 % и уменьшение средней задержки обработки с 500,92 до 175,96 мс.

Ключевые слова: голосовое управление ПК; локальная обработка речи; ASR; NLU; командная грамматика; безопасное исполнение.

Финансирование: Исследования проводились в рамках НИР «Исследование и разработка методов интеллектуального поиска в информационных системах с семантическим ядром (FREN-2026-0004)» (регистрационный номер 126022618063-5).

The relevance of the study is determined by the need for local PC voice control without transferring speech data to external services. The aim of the work is to develop and evaluate a hybrid architecture of a Russian-language voice assistant. The system combines Vosk-based automatic speech recognition with grammar constraints, a natural language understanding model for command interpretation, a rule-based module, and a controlled execution layer. An experiment on 1053 audio samples showed that joint intent–parameter accuracy increased from 83.67% to 96.58%, while the word error rate decreased from 20.05% to 12.56% and mean processing latency decreased from 500.92 to 175.96 ms.

Keywords: PC voice control; on-device speech processing; ASR; NLU; command grammar; safe execution.

Funding: The research was carried out as part of the R&D project 'Research and Development of Methods for Intelligent Search in Information Systems with a Semantic Core (FREN-2026-0004)' (registration number 126022618063-5).

Введение

Системы голосового взаимодействия с ПК являются важным направлением развития интеллектуальных пользовательских интерфейсов. В отличие от диктовки и диалоговых ассистентов, результатом обработки команды здесь является не текстовая транскрипция или ответ системы, а действие в программной среде. Поэтому качество голосового управления определяется не только точностью распознавания речи, но и корректностью интерпретации намерений пользователя (интента), извлечения параметров, проверки допустимости и выполнения действия.

Для локального управления ПК особенно важны автономность, малая задержка, воспроизводимость поведения и отсутствие передачи речевых данных во внешние сервисы. В русскоязычном сценарии задача усложняется морфологической вариативностью, разговорными формулировками, смешением русских и заимствованных названий приложений, а также неоднородностью командной лексики. При этом в открытых материалах по близким решениям недостаточно представлены воспроизводимые оценки полного командного конвейера: от автоматического распознавания речи (*Automatic Speech Recognition, ASR*) и понимания естественного языка (*Natural Language Understanding, NLU*) до контроля допустимости и исполнения действия. В связи с этим сохраняется научно-техническая задача построения компактной локальной архитектуры, в которой автоматическое распознавание речи, NLU-интерпретация, маршрутизация команды и контролируемое исполнение рассматриваются как единый конвейер. Такая архитектура должна обеспечивать корректность конечной команды и её параметров, отказ от неподдерживаемых или потенциально опасных действий, а также малую задержку обработки.

Цель работы состоит в разработке и экспериментальной оценке гибридной архитектуры автономного русскоязычного голосового ассистента для локального управления ПК. Для достижения цели решались следующие задачи: разработка локального конвейера ASR–NLU–исполнение; формирование командной грамматики и набора поддерживаемых интентов; реализация механизма контролируемого исполнения действий; подготовка корпуса командных и негативных сценариев; сравнение режимов ASR с грамматическим ограничением и без него; сопоставление разработанного ассистента с аналогами по методу серого ящика (gray-box).

Научная новизна работы заключается в экспериментальном обосновании гибридной архитектуры, в которой автоматическое распознавание речи с грамматическим ограничением сочетается с NLU-моделью определения интента и параметров, снижая зависимость системы от строго заданных формулировок команд. Показано, что такое сочетание повышает не только качество речевой гипотезы, но и совместную точность интента и параметров: с 83,67 до 96,58 %. Техническая новизна состоит в реализации автономного Windows-конвейера, где локальное ASR, командная грамматика, модель NLU в формате Open Neural Network Exchange (ONNX), модуль правил, маршрутизатор интентов и слой контроля исполнения объединены в систему голосового управления ПК.

Современное состояние систем голосового взаимодействия и управления

Современные голосовые интерфейсы применяются в пользовательских, производственных, сервисных и ассистивных системах. В работах [1–5] рассматриваются вопросы удобства, пользовательского опыта, проактивного поведения, прикладного

использования и этических рисков голосовых ассистентов. Эти исследования подтверждают значимость голосового взаимодействия, однако преимущественно относятся к широкому классу диалоговых и ассистивных решений, тогда как локальное голосовое управление ПК требует отдельного рассмотрения. Наиболее близкими к данной задаче являются исследования локальных, встраиваемых и выполняемых непосредственно на устройстве речевых систем [6–10]. Они показывают реализуемость автономного ASR и понимания устной речи непосредственно на устройстве (Spoken Language Understanding, SLU), но не полностью закрывают задачу русскоязычного командного конвейера управления ПК, где необходимо совместно учитывать распознавание речи, интерпретацию интента, извлечение параметров, маршрутизацию команды и контроль допустимости исполнения.

Для построения компонентов понимания естественного языка важны русскоязычные ресурсы и исследования в области обработки естественного языка (Natural Language Processing, NLP): RussianSuperGLUE [11], ruMTEB [12], работы по архитектурам NLP-систем, языковому моделированию, нормализации, адаптации и коррекции текстов [13–18]. Методическую основу интерпретации команд составляют исследования совместного определения интента и заполнения слотов [19], [20]. Однако в задаче управления ПК распознанный текст является только промежуточным результатом, поэтому итоговое качество должно оцениваться не только по доле пословных ошибок (Word Error Rate, WER) и доле посимвольных ошибок (Character Error Rate, CER), но и по корректности интента, параметров и готовности действия к исполнению. Наряду с этим для локального управления ПК важны исследования коротких речевых команд, обнаружения ключевых слов и ограниченно-словарного распознавания [21–25]. Они подтверждают эффективность работы с конечным набором команд, но чаще оценивают распознавание команды или ключевого слова, а не полный путь от аудиосигнала до безопасного действия в операционной среде.

Практические решения голосового управления ПК представлены встроенными средствами операционных систем, системами диктовки, инструментами пользовательских команд, открытыми голосовыми платформами и настольными ассистентами, включая Voice Access, Dragon, VoiceAttack, Rhasspy, OpenVoiceOS, Laitis, Uvoxus и DecisionsAI. Для таких решений характерна привязка к заранее заданным командам, сценариям или пользовательским правилам, что ограничивает устойчивость к свободным переформулировкам. Поэтому актуален переход от распознавания фиксированных команд к NLU-интерпретации интента и параметров [26].

Таким образом, сохраняется задача разработки компактной локальной архитектуры русскоязычного голосового управления ПК, в которой ASR с грамматическим ограничением, NLU и слой контролируемого исполнения образуют единый командный конвейер. Его оценка должна учитывать не только качество транскрипции, но и корректность интента и параметров, допустимость действия, задержку и безопасность исполнения команды.

Материалы и методы исследования

Исследование выполнено как прикладное архитектурно-экспериментальное и направлено на проверку локального русскоязычного конвейера голосового управления ПК. Методическую основу составили анализ работ по локальному распознаванию речи, интерпретации команд и контролируемому исполнению действий, а также разработка и экспериментальная оценка собственного прототипа. Программно-

технической базой послужил автономный голосовой ассистент для Windows 10/11, реализованный на платформе .NET 8. Система функционирует локально, без обращения к внешним службам, и включает подсистемы ASR, NLU-интерпретации, маршрутизации команды, проверки допустимости действия и исполнительный слой. Материалами исследования служили набор русскоязычных команд управления ПК, конфигурация интентов и параметров, аудиозаписи в формате WAV для контролируемого воспроизведения, а также обучающая, валидационная и тестовая выборки, использованные для подготовки, настройки и проверки подсистемы интерпретации команд. Командное пространство охватывало запуск приложений, переключение окон, навигацию, регулировку системных параметров и служебные операции. Экспериментальный аудиобанк включал записи трёх дикторов: двух мужчин и одной женщины; для оценки устойчивости дополнительно использовались варианты аудио с программной деградацией качества.

Распознавание речи реализовано на базе локальной модели Vosk vosk-model-small-ru-0.22 с автоматически формируемой командной грамматикой, включающей команды, варианты названий приложений, русские числительные, слово активации и дополнительные пользовательские слова. После ASR выполнялась нормализация распознанного текста. Интерпретация команд осуществлялась гибридно: модели семейства ruBERT в формате ONNX использовались для определения интентов и извлечения параметров, а часть коротких и частотных команд обрабатывалась модулем правил. Исполнение отделялось от интерпретации слоем контроля, включающим разрешённый список приложений и подтверждение потенциально опасных операций. Экспериментальная проверка проводилась в интерактивном режиме с микрофонным вводом и в режиме пакетного тестирования. Пакетный режим предусматривал прямой ввод текста, распознавание WAV-файлов и сравнение работы системы при включённом и отключённом грамматическом ограничении ASR. Качество оценивалось по точности определения интента (Intent Accuracy), точности параметров (Parameter Accuracy) и совместной точности интента и параметров (Joint Accuracy). Эти показатели отражали соответственно корректность интента, корректность параметров и их одновременное совпадение с эталоном. Дополнительно рассчитывались WER и CER для оценки речевой гипотезы, полная задержка обработки и пропускная способность.

Архитектура и программная реализация локального голосового ассистента для ПК

При проектировании локального голосового ассистента выбрана модульная архитектура, в которой обработка пользовательской команды представлена как последовательный конвейер: ввод и подготовка аудиосигнала, выделение речевого фрагмента, автоматическое распознавание речи, нормализация текста, интерпретация интента и параметров, маршрутизация действия, проверка допустимости и выполнение операции в среде Windows. Такое разделение соответствует практике построения прикладных систем понимания устной речи, где этапы ASR, определения интента и извлечения параметров реализуются раздельно, но функционируют как единый командный контур [19], [20]. Общая схема архитектуры представлена в соответствии с рисунком 1.

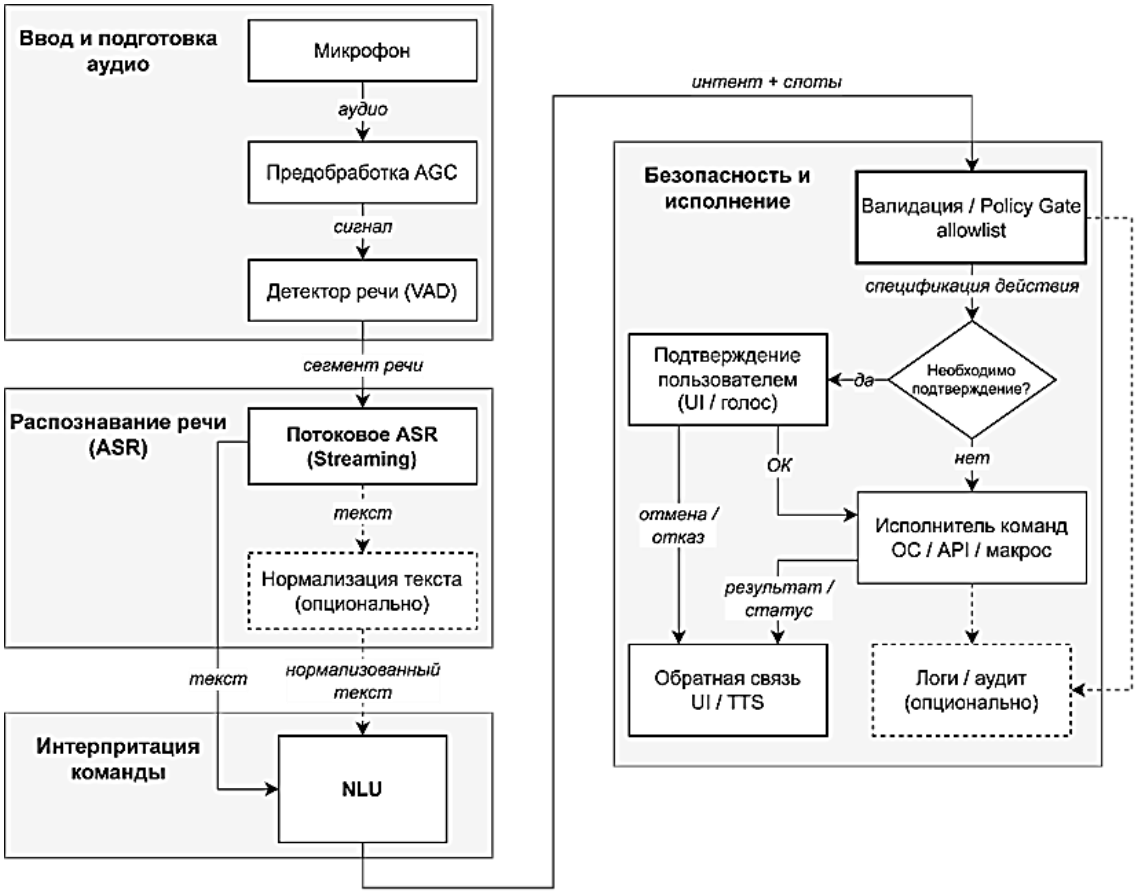


Рисунок 1 – Общая архитектура ассистента

В интерактивном режиме входной аудиосигнал поступает с микрофона, а в режиме пакетного тестирования из WAV-файлов. На первом этапе выполняется предварительная стабилизация сигнала. На схеме этот блок обозначен как автоматическая регулировка усиления (Automatic Gain Control, AGC), поскольку в типовой архитектуре голосовых систем он отвечает за выравнивание уровня аудиопотока перед выделением речи и распознаванием. В разработанном прототипе соответствующие функции реализованы через калибровку микрофонного ввода, оценку уровня фонового шума, шумоподавление и адаптивный анализ энергии аудиокадров. Поэтому AGC в данной архитектуре рассматривается как функциональный блок предобработки, обеспечивающий более устойчивый вход для последующих этапов.

После предварительной обработки применяется детектор речевой активности (Voice Activity Detection, VAD). Он отделяет речевые фрагменты от пауз и фонового шума, определяет границы команды и передаёт в распознавание только релевантный сегмент аудиопотока. В интерактивном режиме дополнительно используется активация по ключевому слову, благодаря чему ассистент переходит от ожидания к записи команды только после обращения пользователя. Это снижает вероятность случайной обработки фоновых звуков как управляющих команд.

Распознавание речи выполняется локально на базе Vosk vosk-model-small-ru-0.22. Для командного режима автоматически формируется грамматика распознавания, ограничивающая пространство поиска словами и фразами предметной области. В неё входят служебные слова, слово активации, шаблоны команд, варианты названий

приложений и папок из разрешённых списков, русские числительные и дополнительные пользовательские фразы. Благодаря этому модуль ASR ориентируется не на произвольную диктовку, а на ограниченное командное пространство локального управления ПК.

После получения речевой гипотезы выполняется нормализация текста: удаляются лишние элементы, унифицируются варианты написания, сглаживаются типичные ошибки распознавания и разговорные формулировки. Нормализованный текст передаётся в блок интерпретации команды. Интерпретация построена по гибридному принципу: модели семейства ruBERT в формате ONNX определяют интент и извлекают параметры, а модуль правил обрабатывает короткие, частотные и однозначные команды. Такое сочетание позволяет сохранять устойчивость к свободным переформулировкам и одновременно повышать предсказуемость обработки простых управляющих фраз.

Результатом интерпретации является структурированное представление команды, включающее интент, параметры, уровень уверенности и служебные диагностические признаки. Далее маршрутизатор сопоставляет интент с конкретным обработчиком действия и передаёт команду в слой контроля исполнения. Этот слой включает валидатор, шлюз политики допустимости, разрешённые списки приложений и операций, а также механизм подтверждения пользователя. Перед выполнением проверяются обязательные параметры, допустимые диапазоны значений и принадлежность объекта действия к разрешённому списку.

Потенциально опасные операции, включая выключение и перезагрузку компьютера, переход в спящий режим, завершение сеанса, закрытие активного окна и принудительное завершение процесса, не выполняются сразу. Они переводятся в состояние ожидания подтверждения и исполняются только после отдельной команды пользователя. Исполнительный слой реализует разрешённые действия средствами Windows-автоматизации: запуск приложений и ссылок, открытие папок, управление окнами, отправку клавиатурных и навигационных событий, регулировку громкости, управление яркостью, медиа-команды и системные операции.

Таким образом, разработанная архитектура объединяет предобработку аудио, VAD-сегментацию, локальное ASR с командной грамматикой, нормализацию текста, гибридную интерпретацию команд, маршрутизацию интентов, проверку допустимости и контролируемое исполнение Windows-действий. Это позволяет рассматривать голосовое управление ПК как полный локальный конвейер преобразования речевой команды в проверенное и допустимое действие. Такое построение снижает зависимость системы от точного совпадения пользовательской фразы с шаблоном и одновременно ограничивает риск нежелательного воздействия на операционную среду.

Экспериментальная оценка разработанного ассистента

Экспериментальная оценка проводилась для проверки качества локального русскоязычного голосового управления ПК и определения влияния грамматического ограничения ASR на точность, задержку и устойчивость обработки команд. Объектом испытаний являлся разработанный прототип локального голосового ассистента: автономная Windows-система с модулем ASR на базе Vosk, моделью NLU в формате ONNX, обработкой на основе правил, маршрутизацией интентов и параметров, слоем контроля исполнения и исполнительным контуром Windows-автоматизации. В тестовой конфигурации использовались модель vosk-model-small-ru-0.22, NLU-модель

distilrubert_small_cased_conversational_contractmax и гибридный режим интерпретации команд. Основной эксперимент выполнялся в пакетном режиме на корпусе из 1053 аудиопримеров в формате WAV: 972 позитивных командных примера и 81 негативный сценарий, для которого корректным результатом являлся отказ от выполнения действия. Аудиоданные включали три профиля сигнала: исходные записи без программной деградации качества (clean), записи с умеренной программной деградацией микрофонного сигнала (bad_mic_light) и записи с выраженной программной деградацией микрофонного сигнала (bad_mic_strong) – по 351 записи в каждом профиле. В записи аудиобанка участвовали три диктора: два мужчины и одна женщина. Измерения выполнялись на ПК с AMD Ryzen 5 5600, 32 ГБ ОЗУ и NVIDIA GeForce RTX 5070.

Сравнивались две конфигурации: режим с грамматическим ограничением распознавания (grammar_on), где ASR использует грамматическую подсказку и ограничивает пространство распознавания командной лексикой, и режим без грамматического ограничения (grammar_off), где распознавание выполняется без такого ограничения. Качество оценивалось по точности определения интента, точности параметров, совместной точности интента и параметров, доле пословных и посимвольных ошибок, задержке обработки и пропускной способности. Сводные результаты приведены в таблице 1.

Таблица 1 – Качество и быстродействие разработанной системы в режимах grammar_on и grammar_off

Метрика	grammar_on	grammar_off	Эффект grammar_on
Точность интента, %	97,06	84,24	+12,82 п.п.
Точность параметров, %	96,91	80,86	+16,05 п.п.
Совместная точность, %	96,58	83,67	+12,91 п.п.
WER, %	12,56	20,05	−7,49 п.п.
CER, %	6,34	8,20	−1,86 п.п.
Средняя задержка обработки, мс	175,96	500,92	−324,96 мс
Медианная задержка обработки, мс	59,00	496,00	−437,00 мс
Пропускная способность, прим./с	5,68	2,00	+3,68

Результаты в таблице 1 показывают, что грамматическое ограничение ASR улучшает не только речевую гипотезу, но и итоговую интерпретацию команды. В режиме grammar_on совместная точность интента и параметров выросла с 83,67 до 96,58 %, доля пословных ошибок снизилась с 20,05 до 12,56 %, а доля посимвольных ошибок с 8,20 до 6,34 %. Наибольший прирост получен по точности параметров: с 80,86 до 96,91 %, что особенно важно для команд, где ошибка параметра может привести к выбору неверного приложения, окна или значения системной настройки. Для наглядного сопоставления основных метрик качества результаты режимов grammar_on и grammar_off представлены в соответствии с рисунком 2.

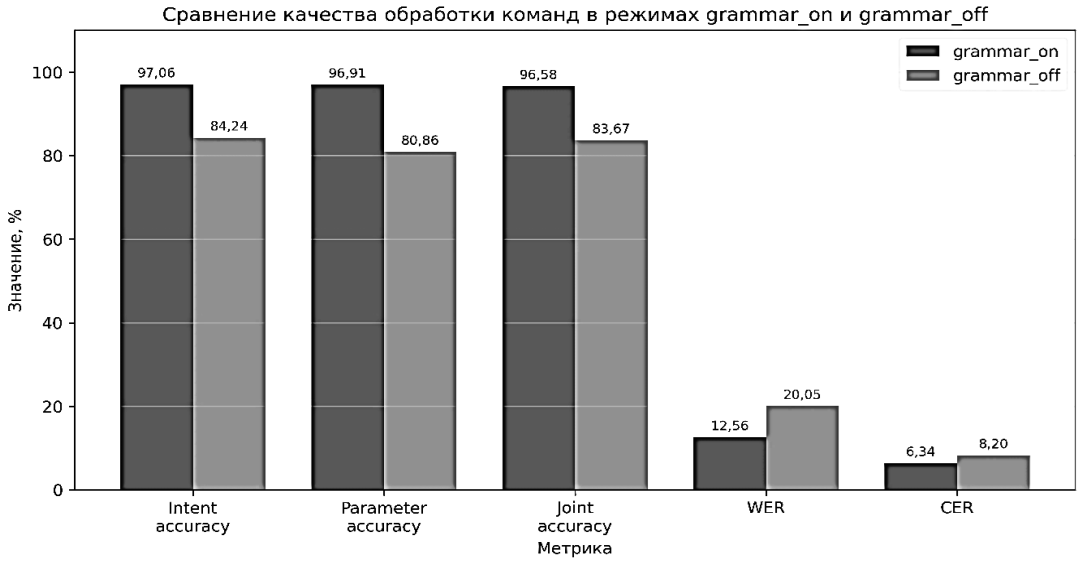


Рисунок 2 – Влияние грамматического ограничения ASR на качество обработки команд

По временным характеристикам режим с грамматическим ограничением распознавания также обеспечил лучшие результаты: средняя задержка обработки снизилась с 500,92 до 175,96 мс, медианная с 496,00 до 59,00 мс, а пропускная способность увеличилась с 2,00 до 5,68 примера в секунду. При этом задержка NLU в обоих режимах оставалась малой и не превышала 10 мс, что подтверждает применимость модели NLU в локальном командном конвейере. Дополнительное профилирование показало меньшую вычислительную нагрузку в режиме с грамматическим ограничением, поэтому данный режим является предпочтительным не только по качеству, но и по вычислительной эффективности. Для оценки практической конкурентоспособности было выполнено сопоставление разработанной системы с Laitis и Uvoxus по методу серого ящика. Для сторонних систем использовались доступные инструментированные события, безопасные действия-заглушки и внешние наблюдатели, поэтому результаты следует рассматривать как прикладное сопоставление поведения систем на общем корпусе, а не как строго тождественное сравнение внутренних этапов по методу белого ящика (white-box). Сводные результаты сопоставления приведены в таблице 2, а соотношение командной точности и медианной задержки показано в соответствии с рисунком 3.

Полученные данные показывают, что разработанная система в режиме с грамматическим ограничением достигает более высокой итоговой командной точности, чем рассмотренные аналоги: совместная точность интента и параметров составила 96,58 % против 80,14 % по метрике успешности положительных действий (Positive Action Success) у Laitis и 83,33 % по метрике точности выбора инструмента (Tool Accuracy) у Uvoxus. При этом Laitis имеет меньшую долю пословных ошибок: 10,67 % против 12,56 % у разработанного ассистента. Следовательно, для систем голосового управления ПК низкая пословная ошибка распознавания не является достаточным критерием качества: конечный результат определяется не только ASR, но и способностью распознавать интент и параметры при вариативных формулировках команды, а также проверкой допустимости и маршрутизацией действия. По сочетанию командной точности и медианной задержки наиболее благоприятное положение занимает режим с грамматическим ограничением. Медианная задержка

разработанного ассистента составила 59,00 мс, тогда как для Uvoxus она равна 914,33 мс, а для Laitis 2169,00 мс. При интерпретации этих значений необходимо учитывать различия в границах измерения задержки, поэтому корректный вывод состоит не в прямом абсолютном превосходстве по всем сценариям, а в том, что предложенная архитектура обеспечивает низкую задержку в ограниченном локальном командном пространстве.

Отдельно проверялись негативные сценарии. В режиме grammar_on все 81 неподдерживаемое или нерелевантное высказывание было отнесено к классу неизвестной команды (unknown), а доля ложных исполнений составила 0,0 %. Следовательно, повышение точности на позитивных командах не сопровождалось ростом ложных исполнений, что подтверждает корректное поведение системы при отказе от выполнения действия.

Таблица 2 – Сопоставление разработанной системы с аналогами по методу серого ящика

Система	Командная метрика	WER, %	Негативные сценарии	Задержка, мс: средняя / медианная
Разработанная система, grammar_on	Совместная точность интента и параметров 96,58 %	12,56	81/81; ложные исполнения = 0,0 %	175,96 / 59,00
Разработанная система, grammar_off	Совместная точность интента и параметров 83,67 %	20,05	—	500,92 / 496,00
Laitis	Успешность положительных действий 80,14 %	10,67	Отклонение негативных сценариев 100,00 %	2213,15 / 2169,00
Uvoxus	Точность выбора инструмента 83,33 %	18,94	Отклонение негативных сценариев 100,00 %	— / 914,33

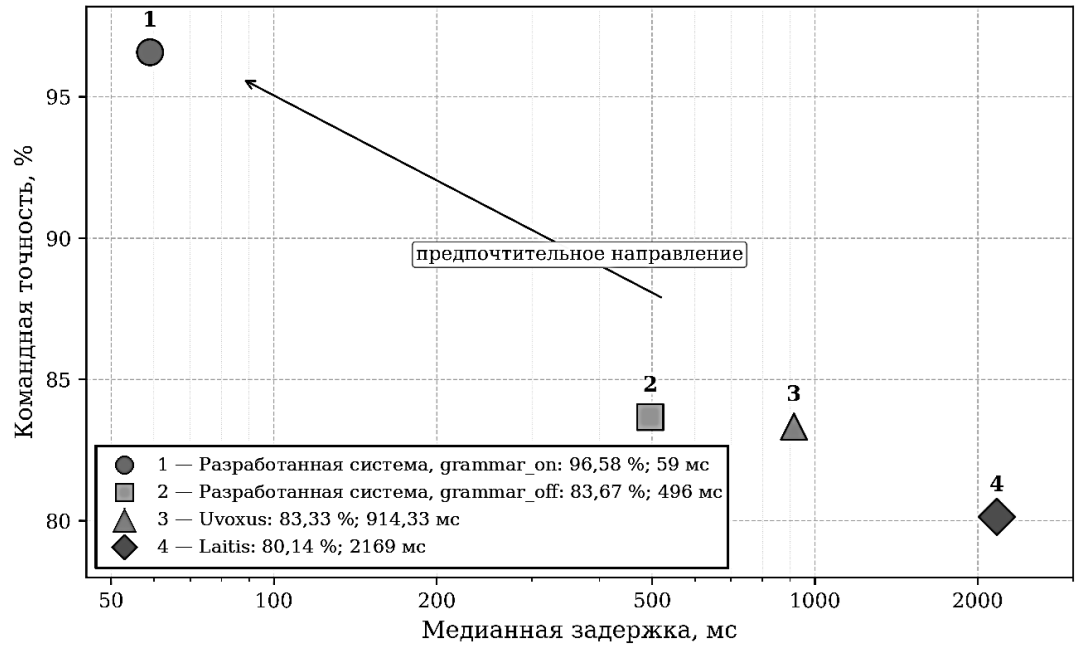


Рисунок 3 – Сопоставление голосовых ассистентов по командной точности и медианной задержке

Таким образом, экспериментальная оценка подтверждает эффективность предложенной архитектуры для локального русскоязычного голосового управления ПК. Режим с грамматическим ограничением распознавания (grammar_on) обеспечивает более высокую точность определения интента и параметров, меньшую долю ошибок распознавания речи, более низкую задержку, большую пропускную способность и корректный отказ от выполнения действий на негативных сценариях. Сопоставление с Laitis и Uvoxus по методу серого ящика показывает конкурентоспособность разработанного ассистента, но должно интерпретироваться с учётом различий в командных метриках, архитектуре систем и границах измерения задержки.

Заключение

В работе предложена и экспериментально проверена гибридная архитектура автономного русскоязычного голосового ассистента для локального управления ПК. Система реализует командный конвейер «речь → распознавание → интерпретация → проверка допустимости → выполнение действия» и объединяет локальное ASR на базе Vosk, грамматическое ограничение командного пространства, модель NLU в формате ONNX, обработку на основе правил и слой контролируемого исполнения. Эксперимент на 1053 аудиопримерах в формате WAV показал, что для рассматриваемой задачи предпочтительным является режим с грамматическим ограничением. При его использовании точность определения интента составила 97,06 %, точность параметров 96,91 %, а совместная точность интента и параметров 96,58 %. По сравнению с режимом без грамматического ограничения совместная точность выросла с 83,67 до 96,58 %, доля пословных ошибок снизилась с 20,05 до 12,56 %, а средняя задержка уменьшилась с 500,92 до 175,96 мс. Это подтверждает, что грамматическое ограничение ASR повышает не только качество речевой гипотезы, но и устойчивость последующей интерпретации команды.

Сопоставление с Laitis и Uvoxus по методу серого ящика показало конкурентоспособность разработанного ассистента по итоговой командной точности и задержке обработки. При этом сравнение следует интерпретировать с учётом различий в архитектуре систем, командных метриках и границах измерения задержки. Практическая значимость работы состоит в возможности использования предложенной архитектуры как основы персонального автономного ассистента, не передающего речевые данные во внешние службы. Дальнейшее развитие системы целесообразно направить на расширение корпуса негативных сценариев, повышение устойчивости к деградации микрофонного сигнала, улучшение обработки навигационных и оконных команд, а также повторные испытания на различных аппаратных конфигурациях.

Список источников

1. Deshmukh A. M. User Experience and Usability of Voice User Interfaces: A Systematic Literature Review / A. M. Deshmukh, R. Chalmeta // Information. 2024. Vol. 15, No. 9. Art. 579. DOI: 10.3390/info15090579.
2. A review of digital assistants in production and logistics: applications, benefits, and challenges / T. Zheng, E. H. Grosse, S. Morana, C. H. Glock // International Journal of Production Research. 2024. Vol. 62, No. 21. P. 8022–8048. DOI: 10.1080/00207543.2024.2330631.
3. Proactive behavior in voice assistants: A systematic review and conceptual model / C. Bérubé, M. Nißen, R. Vinay [et al.] // Computers in Human Behavior Reports. 2024. Vol. 14. Art. 100411. DOI: 10.1016/j.chbr.2024.100411.
4. Balan C. Chatbots and Voice Assistants: Digital Transformers of the Company-Customer Interface – A Systematic Review of the Business Research Literature / C. Balan // Journal of Theoretical and Applied Electronic Commerce Research. 2023. Vol. 18, No. 2. P. 995–1019. DOI: 10.3390/jtaer18020051.

5. A Systematic Review of Ethical Concerns with Voice Assistants / W. Seymour, X. Zhan, M. Coté, J. M. Such // Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. 2023. P. 131–145. DOI: 10.1145/3600211.3604679.
6. Lazzaroni L. An embedded end-to-end voice assistant / L. Lazzaroni, F. Bellotti, R. Berta // Engineering Applications of Artificial Intelligence. 2024. Vol. 136. Art. 108998. DOI: 10.1016/j.engappai.2024.108998.
7. Di Leo S. Real-Time Speech-to-Text on Edge: A Prototype System for Ultra-Low Latency Communication with AI-Powered NLP/ Di Leo S., De Cicco L., S. Mascolo // Information. 2025. Vol. 16, No. 8. Art. 685. DOI: 10.3390/info16080685.
8. Gondi S., Pratap V. Performance Evaluation of Offline Speech Recognition on Edge Devices // Electronics. 2021. Vol. 10, No. 21. Art. 2697. DOI: 10.3390/electronics10212697.
9. Deliberation Model for On-Device Spoken Language Understanding / D. Le, A. Shrivastava, P. D. Tomasello [et al.] // Proceedings of Interspeech 2022. 2022. P. 3468–3472. DOI: 10.21437/Interspeech.2022-10800.
10. Do T., Nguyen M.-P., Nguyen L.-M. PruneSLU: Efficient On-device Spoken Language Understanding through Vocabulary and Structural Pruning // Proceedings of Interspeech 2025. 2025. P. 1738–1742. DOI: 10.21437/Interspeech.2025-1893.
11. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / T. Shavrina, A. Fenogenova, A. Emelyanov [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020. P. 4717–4726. DOI: 10.18653/v1/2020.emnlp-main.381.
12. The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design / A. Snegirev, M. Tikhonova, A. Maksimova [et al.] // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025. P. 236–254. DOI: 10.18653/v1/2025.naacl-long.12.
13. Пикалёв Я. С. Обзор архитектур систем интеллектуальной обработки естественно-языковых текстов / Я.С. Пикалев // Проблемы искусственного интеллекта. 2020. № 4(19). С. 45–68.
14. Ермоленко Т. В. Адаптация нейросетевой модели ALBERT для задачи языкового моделирования / Т.В. Ермоленко, Я. С. Пикалёв // Проблемы искусственного интеллекта. 2020. № 3(18). С. 111–122.
15. Пикалёв Я. С. Разработка системы нормализации текстовых корпусов / Я.С. Пикалев // Проблемы искусственного интеллекта. 2022. № 2(25). С. 64–78.
16. Большакова С. А. Система автоматической адаптации русскоязычных текстов и ее практическая значимость / С.А. Большакова // Проблемы искусственного интеллекта. 2024. № 3(34). С. 45–54. DOI: 10.24412/2413-7383-2024-3-45-54.
17. Елисеев В. О. Разработка интеллектуального ассистента для образовательной организации / В. О. Елисеев, В. И. Бондаренко, А. Ю. Максимова // Проблемы искусственного интеллекта. 2025. № 4(39). С. 36–48. DOI: 10.24412/2413-7383-2025-4-39-36-48.
18. Шелепов В. Ю. О текстовой коррекции и ее применении при распознавании речи / В.Ю. Шелепов, А. В. Ниценко // Проблемы искусственного интеллекта. 2017. № 4(7). С. 50–56.
19. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding / H. Weld, X. Huang, S. Long, J. Poon, S. C. Han // ACM Computing Surveys. 2022. Vol. 55, No. 8. Art. 156. DOI: 10.1145/3547138.
20. Firdaus M. Multitask learning for multilingual intent detection and slot filling in dialogue systems / M. Firdaus, A. Ekbal, E. Cambria // Information Fusion. 2023. Vol. 91. P. 299–315. DOI: 10.1016/j.inffus.2022.09.029.
21. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford, J. W. Kim, T. Xu [et al.] // Proceedings of the 40th International Conference on Machine Learning. 2023. Vol. 202. P. 28492–28518.
22. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition / P. Warden // arXiv. 2018. arXiv:1804.03209. DOI: 10.48550/arXiv.1804.03209.
23. Rusci M. Few-Shot Open-Set Learning for On-Device Customization of KeyWord Spotting Systems / M. Rusci, T. Tuytelaars // Proceedings of Interspeech 2023. 2023. P. 2768–2772. DOI: 10.21437/Interspeech.2023-1904.
24. Robust Keyword Spotting for Noisy Environments by Leveraging Speech Enhancement and Speech Presence Probability / C. Yang, Y. M. Saidutta, R. S. Srinivasa [et al.] // Proceedings of Interspeech 2023. 2023. P. 1638–1642. DOI: 10.21437/Interspeech.2023-2222.
25. Robust speech command recognition in challenging industrial environments / Bini S., Carletti V., Saggese A., Vento M. // Computer Communications. 2024. Vol. 228. Art. 107938. DOI: 10.1016/j.comcom.2024.107938.
26. Сидорика М. Д. Разработка автономного голосового интерфейса для управления рабочим столом: от распознавания до семантического анализа / М.Д. Сидорика, Л. А. Лазебная // Сборник трудов ИУСМКМ-2025. Т. 2. Донецк: ДонНТУ, 2025. С. 265.

References

1. Deshmukh A. M. User Experience and Usability of Voice User Interfaces: A Systematic Literature Review / A. M. Deshmukh, R. Chalmeta // Information. – 2024. – Vol. 15, No. 9. – Art. 579. – DOI: 10.3390/info15090579.
2. A review of digital assistants in production and logistics: applications, benefits, and challenges. International Journal of Production Research / Zheng T., Grosse E. H., Morana S., Glock C. H. – 2024. – Vol. 62, No. 21. – P. 8022–8048. – DOI: 10.1080/00207543.2024.2330631.
3. Proactive behavior in voice assistants: A systematic review and conceptual model / Bérubé C., Nißen M., Vinay R. [et al.] // Computers in Human Behavior Reports. – 2024. – Vol. 14. – Art. 100411. – DOI: 10.1016/j.chbr.2024.100411.
4. Balan C. Chatbots and Voice Assistants: Digital Transformers of the Company-Customer Interface – A Systematic Review of the Business Research Literature / C. Balan // Journal of Theoretical and Applied Electronic Commerce Research. – 2023. – Vol. 18, No. 2. – P. 995–1019. – DOI: 10.3390/jtaer18020051.
5. A Systematic Review of Ethical Concerns with Voice Assistants / Seymour W., Zhan X., Coté M., Such J. M. // Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. – 2023. – P. 131–145. – DOI: 10.1145/3600211.3604679.
6. Lazzaroni L. An embedded end-to-end voice assistant / L. Lazzaroni, F. Bellotti, R. Berta // Engineering Applications of Artificial Intelligence. – 2024. – Vol. 136. – Art. 108998. – DOI: 10.1016/j.engappai.2024.108998.
7. Di Leo S. Real-Time Speech-to-Text on Edge: A Prototype System for Ultra-Low Latency Communication with AI-Powered NLP / Di Leo S., De Cicco L., Mascolo S. // Information. – 2025. – Vol. 16, No. 8. – Art. 685. – DOI: 10.3390/info16080685.
8. Gondi S. Performance Evaluation of Offline Speech Recognition on Edge Devices / S. Gondi, V. Pratap // Electronics. – 2021. – Vol. 10, No. 21. – Art. 2697. – DOI: 10.3390/electronics10212697.
9. Deliberation Model for On-Device Spoken Language Understanding / Le D., Shrivastava A., Tomasello P. D. [et al.] // Proceedings of Interspeech 2022. – 2022. – P. 3468–3472. – DOI: 10.21437/Interspeech.2022-10800.
10. Do T. PruneSLU: Efficient On-device Spoken Language Understanding through Vocabulary and Structural Pruning / T. Do, M.-P. Nguyen, L.-M. Nguyen // Proceedings of Interspeech 2025. – 2025. – P. 1738–1742. – DOI: 10.21437/Interspeech.2025-1893.
11. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / Shavrina T., Fenogenova A., Emelyanov A. [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. – 2020. – P. 4717–4726. – DOI: 10.18653/v1/2020.emnlp-main.381.
12. The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design / Snegirev A., Tikhonova M., Maksimova A. [et al.] // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2025. – P. 236–254. – DOI: 10.18653/v1/2025.naacl-long.12.
13. Pikalev Ya. S. Overview of architectures of intelligent natural-language text processing systems / Ya. S. Pikalev // Problems of Artificial Intelligence. – 2020. – No. 4(19). – P. 45–68.
14. Ermolenko T. V. Adaptation of the ALBERT neural network model for the language modeling task / T. V. Ermolenko, Ya. S. Pikalev // Problems of Artificial Intelligence. – 2020. – No. 3(18). – P. 111–122.
15. Pikalev Ya. S. Development of a text corpus normalization system / Ya. S. Pikalev // Problems of Artificial Intelligence. – 2022. – No. 2(25). – P. 64–78.
16. Bolshakova S. A. System for automatic adaptation of Russian-language texts and its practical significance / S. A. Bolshakova // Problems of Artificial Intelligence. – 2024. – No. 3(34). – P. 45–54. – DOI: 10.24412/2413-7383-2024-3-45-54.
17. Eliseev V. O. Development of an intelligent assistant for an educational organization / V. O. Eliseev, V. I. Bondarenko, A. Yu. Maksimova // Problems of Artificial Intelligence. – 2025. – No. 4(39). – P. 36–48. – DOI: 10.24412/2413-7383-2025-4-39-36-48.
18. Shelepov V. Yu. On text correction and its application in speech recognition / V. Yu. Shelepov, A. V. Nitsenko // Problems of Artificial Intelligence. – 2017. – No. 4(7). – P. 50–56.
19. A Survey of Joint Intent Detection and Slot Filling Models in Natural Language Understanding / Weld H., Huang X., Long S., Poon J., Han S. C. // ACM Computing Surveys. – 2022. – Vol. 55, No. 8. – Art. 156. – DOI: 10.1145/3547138.
20. Firdaus M. Multitask learning for multilingual intent detection and slot filling in dialogue systems / M. Firdaus, A. Ekbal, E. Cambria // Information Fusion. – 2023. – Vol. 91. – P. 299–315. – DOI: 10.1016/j.inffus.2022.09.029.

21. Robust Speech Recognition via Large-Scale Weak Supervision / Radford A., Kim J. W., Xu T. [et al.] // Proceedings of the 40th International Conference on Machine Learning. – 2023. – Vol. 202. – P. 28492–28518.
22. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition / P. Warden // arXiv. – 2018. – arXiv:1804.03209. – DOI: 10.48550/arXiv.1804.03209.
23. Rusci M. Few-Shot Open-Set Learning for On-Device Customization of KeyWord Spotting Systems / M. Rusci T. Tuytelaars // Proceedings of Interspeech 2023. – 2023. – P. 2768–2772. – DOI: 10.21437/Interspeech.2023-1904.
24. Robust Keyword Spotting for Noisy Environments by Leveraging Speech Enhancement and Speech Presence Probability / Yang C., Saidutta Y. M., Srinivasa R. S. [et al.] // Proceedings of Interspeech 2023. – 2023. – P. 1638–1642. – DOI: 10.21437/Interspeech.2023-2222.
25. Robust speech command recognition in challenging industrial environments / Bini S., Carletti V., Saggese A., Vento M. // Computer Communications. – 2024. – Vol. 228. – Art. 107938. – DOI: 10.1016/j.comcom.2024.107938.
26. Sidorika M. D. Development of an autonomous voice interface for desktop control: from recognition to semantic analysis / M. D. Sidorika, L. A. Lazebnaya // Proceedings of IUSMKM-2025. – Vol. 2. – Donetsk: DonNTU, 2025. – 265 p.

RESUME

Sidorika M. D., Nitsenko A.V.

Hybrid Architecture of an Autonomous Voice Assistant for Local PC Control

The development of local PC voice control systems is relevant due to the need to process speech commands without transferring user data to external services. Such systems must provide low latency, predictable behavior, reliable command interpretation, and protection against unintended actions, since the result is an executable action rather than a transcript.

The paper proposes a hybrid architecture of a Russian-language voice assistant for Windows 10/11. The system combines local Vosk-based ASR with grammar constraints, an NLU model for intent and parameter extraction, rule-based processing, command routing, and controlled execution. The evaluation was carried out in batch mode on 1053 WAV samples, including 972 positive commands and 81 negative scenarios.

The experiment showed that grammar-constrained ASR improves both speech recognition and command interpretation. Joint Accuracy increased from 83.67% to 96.58%, WER decreased from 20.05% to 12.56%, and mean latency decreased from 500.92 to 175.96 ms. NLU latency remained below 10 ms, and all negative scenarios were rejected correctly with a false execution rate of 0.0%.

The results confirm the effectiveness of combining grammar-constrained ASR, NLU-based command interpretation, rule-based processing, and controlled execution for offline PC voice control.

РЕЗЮМЕ

Сидорика М. Д., Ниценко А.В.

Гибридная архитектура автономного голосового ассистента для локального управления ПК

Разработка систем локального голосового управления ПК актуальна из-за необходимости обработки речевых команд без передачи пользовательских данных во внешние сервисы. Такие системы должны обеспечивать малую задержку, предсказуемое поведение, надёжную интерпретацию команд и защиту от нежелательных действий, поскольку результатом обработки является не транскрипция, а выполняемое действие.

В работе предложена гибридная архитектура русскоязычного голосового ассистента для Windows 10/11. Система объединяет локальное автоматическое распознавание речи на базе Vosk с грамматическим ограничением, модель понимания

естественного языка для определения интентов и параметров, обработку на основе правил, маршрутизацию команд и контролируемое исполнение. Оценка проводилась в пакетном режиме на 1053 аудиопримерах в формате WAV, включая 972 позитивные команды и 81 негативный сценарий.

Эксперимент показал, что грамматическое ограничение ASR улучшает распознавание речи и интерпретацию команд. Совместная точность интента и параметров увеличилась с 83,67 до 96,58 %, доля пословных ошибок снизилась с 20,05 до 12,56 %, а средняя задержка уменьшилась с 500,92 до 175,96 мс. Задержка модели понимания естественного языка оставалась ниже 10 мс, все негативные сценарии были корректно отклонены при доле ложных исполнений 0,0 %.

Полученные результаты подтверждают эффективность сочетания грамматически ограниченного ASR, NLU-интерпретации команд, обработки на основе правил и контролируемого исполнения для автономного голосового управления ПК.

Сидорика Максим Денисович – стажёр-исследователь отдела распознавания речевых образов, ФГБНУ «Институт проблем искусственного интеллекта», г. Донецк. *Область научных интересов*: искусственный интеллект, голосовые пользовательские интерфейсы, понимание естественного языка, голосовое управление ПК, интеллектуальные ассистенты, программная автоматизация. E-mail: sidorika.maks@gmail.com. ORCID: 0009-0008-9388-324X. Адрес: 283048, г. Донецк, ул. Артема, д. 118-б.

Ниценко Артём Владимирович – к.т.н., старший научный сотрудник отдела распознавания речевых образов, ФГБНУ «Институт проблем искусственного интеллекта», г. Донецк. *Область научных интересов*: искусственный интеллект, обработка естественного языка, компьютерная лингвистика, распознавание речи. E-mail: nav_box@mail.ru. ORCID: 0009-0008-3411-7107. Адрес: 283048, г. Донецк, ул. Артема, д. 118-б.

Sidorika Maxim Denisovich – Trainee Researcher of Speech Recognition Department, Federal State Budgetary Scientific Institution “Institute of Artificial Intelligence Problems”, 283048, Donetsk, Artema St., 118b, sidorika.maks@gmail.com. Research interests: artificial intelligence, voice user interfaces, natural language understanding, voice control of PCs, intelligent assistants, software automation. ORCID: 0009-0008-9388-324X.

Nitsenko Artyom Vladimirovich – Candidate of Technical Sciences, Senior Researcher of Speech Recognition Department, Federal State Budgetary Scientific Institution “Institute of Artificial Intelligence Problems”, 283048, Donetsk, Artema St., 118b, nav_box@mail.ru Research interests: artificial intelligence, natural language processing, computational linguistics, speech recognition. ORCID: 0009-0008-3411-7107.

Статья поступила в редакцию 30.03.2026;
одобрена после рецензирования 30.04.2026; принята к публикации 02.09.2026
The article was submitted 30.03.2026; approved after reviewing 30.04.2026;
accepted for publication 02.09.2026.