

УДК 004.934.1'1

Особенности реализации системы голосового управления для AR-гарнитуры

Кондрашин Матвей Евгеньевич^(*)

matkondr@yandex.ru

Кузнецов Алексей Станиславович

askuznetsov@bmstu.ru

SPIN-код: 8909-3004

Волкова Мария Валентиновна

mvvvolkova@bmstu.ru

SPIN-код: 5701-0182

МГТУ им. Н.Э. Баумана, Москва, Россия

Аннотация. Рассмотрены различные методы и технологии распознавания голосовых команд для реализации голосового управления AR-гарнитурой. Обоснована выбранная схема алгоритма обработки с применением готовых модулей распознавания. Описаны особенности работы, параметры и функционал разработанной программно-аппаратный комплекс AR-гарнитуры с голосовым управлением на основе микроконтроллера Banana Pi CM4. Проведена оценка точности распознавания команд для разных микрофонов.

Ключевые слова: дополненная реальность, голосовое управление, распознавание речи, нейронные сети

Features of the implementation of the voice control system for AR headsets

Kondrashin Matvey Evgenievich^(*)

matkondr@yandex.ru

Kuznetsov Alexey Stanislavovich

askuznetsov@bmstu.ru

SPIN-code: 8909-3004

Volkova Maria Valentinovna

mvvvolkova@bmstu.ru

SPIN-code: 5701-0182

BMSTU, Moscow, Russia

Abstract. Various methods and technologies for recognizing voice commands for implementing voice control of an AR headset are considered. The chosen scheme of the processing algorithm using ready-made recognition modules is justified. The features of operation, parameters and functionality of the developed AR headset hardware and software complex with voice control based on the Banana Pi CM4 microcontroller are described. The accuracy of command recognition for different microphones was evaluated.

Keywords: augmented reality, voice control, speech recognition, neural networks

Введение. В контексте развития миниатюрных микропроцессорных устройств, оптических модулей отображения и нейросетевых технологий, растет тренд на разработку и внедрение AR-устройств для применения в таких областях как промышленность, добыча полезных ископаемых, медицина, образование,

транспорт и обслуживание техники [1, 2]. Одним из преимуществ наголовных AR гарнитур является возможность использования в режиме «без рук». Это обеспечивается использованием управления жестами и голосовыми командами. Оба метода имеют свои преимущества и области применения, которые могут пересекаться [3].

Управление жестами дает удобную возможность работать со сложными 3D-моделями и чертежами в режиме наложенной реальности. Однако это не самый точный и удобный метод для навигации по меню и работы с простыми интерфейсами, например файлового проводника. Кроме того, в условиях промышленного применения руки пользователя могут быть заняты инструментом, что исключает возможность такого способа управления. Помимо этого, длительное управление гарнитурой жестами требует постоянного нахождения рук в поле зрения камеры, что приводит к напряжению мышц и усталости. Управление голосовыми командами проще в плане программной реализации, нагрузки на процессор и не требует участия рук, что может быть полезно при проведении сложных работ или для людей с инвалидностью.

Основной программный функционал разрабатываемого AR-устройства включает в себя осуществление аудио и видеозвонков, просмотр медиаданных, запись видео и изображений, файловый менеджер. Голосовое управление является оптимальным для данных задач, которые требуют четкие инструкции.

Существует два подхода в обработке голосовых команд. Первый подразумевает распознавание строго по заранее составленному списку команд, что требует от пользователя четкости произношения и запоминания команд. Вторым подход заключается в семантической обработке команд, которые пользователь дает в свободной форме. В статье [4] показано, что второй подход обеспечивает большую точность отклика на команды и помогает обучать пользователей работе с устройством, так как не нужно запоминать список четких команд. Однако такой способ требует большего времени отклика от устройства из-за более сложной реализации.

Материалы и методы. Для демонстрации возможности реализации системы голосового управления AR-гарнитурой использовался микроконтроллер Banana PI CM4. Данный модуль, в отличие от Raspberry PI CM4, имеет более мощный процессор и встроенный нейропроцессор со скоростью 5TOPs. За счет этого распознавание команд происходит быстрее. Ниже приведены характеристики данного устройства:

- габариты платы — 55 × 40 мм;
- процессор — 4 ядра Cortex-A73 , 2 ядра Cortex-A53;
- тактовая частота — 2,0 ГГц;
- оперативная память — 4 Гб;
- поддержка видео — MIPI, HDMI;
- операционная система — Linux Ubuntu или Android.

На рис. 1 приведена схема подключения внешних устройств к модулю AR-гарнитур. Возможно два варианта реализации распознавания голосовых

команд [1]. Первый подразумевает использование облачных сервисов распознавания речи таких как: Google Speech API, Alexa Voice Services, Watson Assistant, Microsoft LUIS [5]. Однако поскольку предполагается автономное использование AR-гарнитуры, такой способ не подходит.

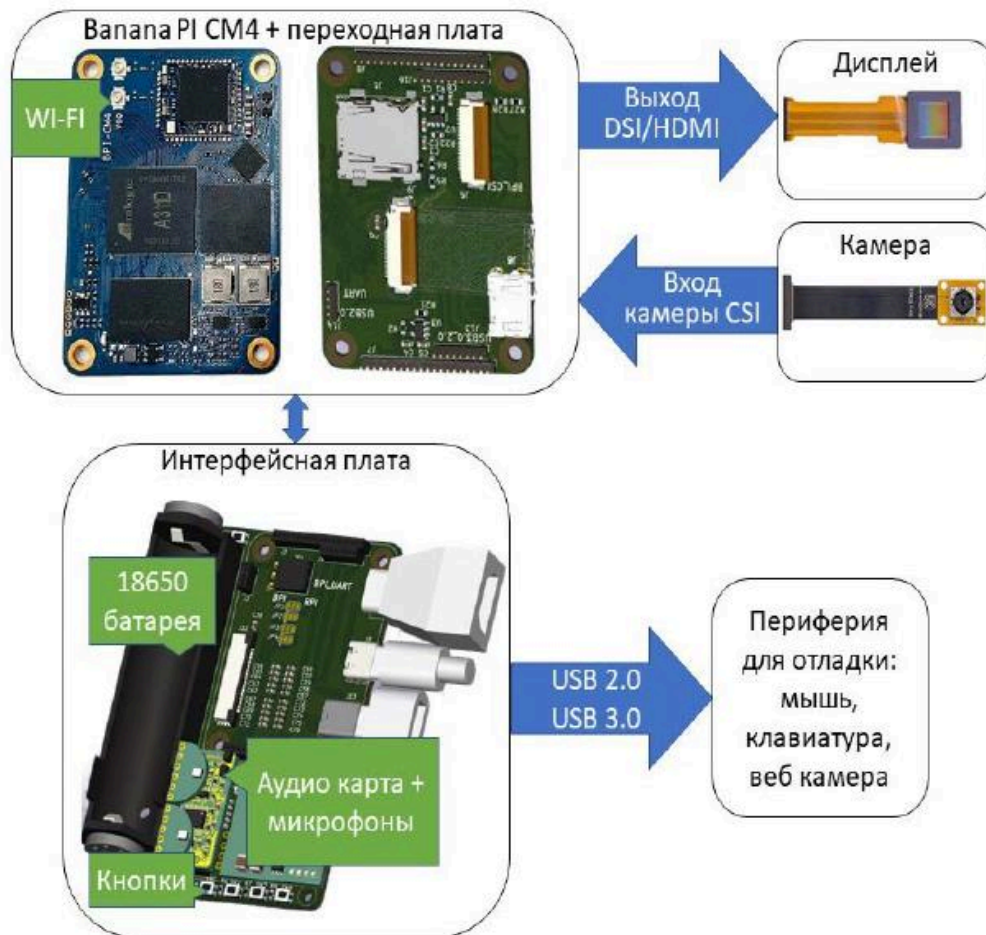


Рис. 1. Схема модуля на основе BPI CM4

Второй вариант заключается в использовании автономной модели распознавания речи. Существует множество популярных библиотек распознавания речи. Ниже приведены некоторые из них в таблице. Real time factor — отношение времени обработки звукового сигнала к его длительности. Как видно из таблицы, по этому параметру самая лучшая модель — Porcupine, которая работает по заранее заданным ключевым словам, что обеспечит лучшую производительность, однако может ограничить функционал. Библиотека Whisper обеспечивает самую высокую точность за счет большого веса и долгой обработки. NVIDIA NeMo также обеспечивает высокую точность при наличии графического вычислительного модуля. Самые оптимальные для реализации на микроконтроллерах за счет малого веса модели, достаточной точности и скорости обработки — модели Silero и Vosk. Окончательный выбор был сделан в пользу Vosk.

Популярные модели распознавания голоса

Название модели	Real time factor	Вес модели, Мб	Точность распознавания
Silero	0,1–0,3	16	Средняя
Whisper	0,5–1,2	461	Высокая
Vosk	0,2–0,5	50	Средняя
NVIDIA NeMo	0,1–0,4 (с GPU)	50	Высокая
Porcupine	< 0,05	< 5	Низкая

Результаты. Для разработки и отладки программного обеспечения использовалась среда разработки Visual Studio Code с подключением к Banana PI по протоколу SSH. Алгоритм распознавания команд состоит из следующих этапов, которые отображены на рис. 2.



Рис. 2. Алгоритм работы модуля распознавания команд

Алгоритм распознавания команд:

- по кнопке запускается модуль распознавания и начинает принимать аудиоданные;
- запустив аудиопоток, модуль распознавания пропускает звуковой сигнал через цифровой фильтр и усилитель;
- зафиксировав всплеск аудиосигнала производится попытка получить текст из сигнала;
- на выходе распознавателя получается текстовый поток отдельными словами или фразами;

- полученный текст проверяется на соответствие текстовой инструкции списку команд по категориям;
- происходит выполнение команд либо в отдельном потоке, либо мгновенно;
- ожидание следующей команды.

В процессе добавления новых команд в список (всего было реализовано 29 команд) для повышения точности распознавания к каждой команде привязывалось несколько вариантов голосовых инструкций. В процессе тестирования работы распознавания выявились наиболее частые ошибочно распознанные голосовые инструкции, которые также привязывались к командам. Ниже приведен список основных команд:

- команды питания;
- команды подключения к сети;
- команды тестирования;
- команды стриминга видео;
- команды навигации по файловому проводнику.

Ниже на рис. 3 приведен пример организации структуры команд вместе с разнообразными вариантами произношения каждой из них, включая варианты в который встречаются ошибки распознавания текста.

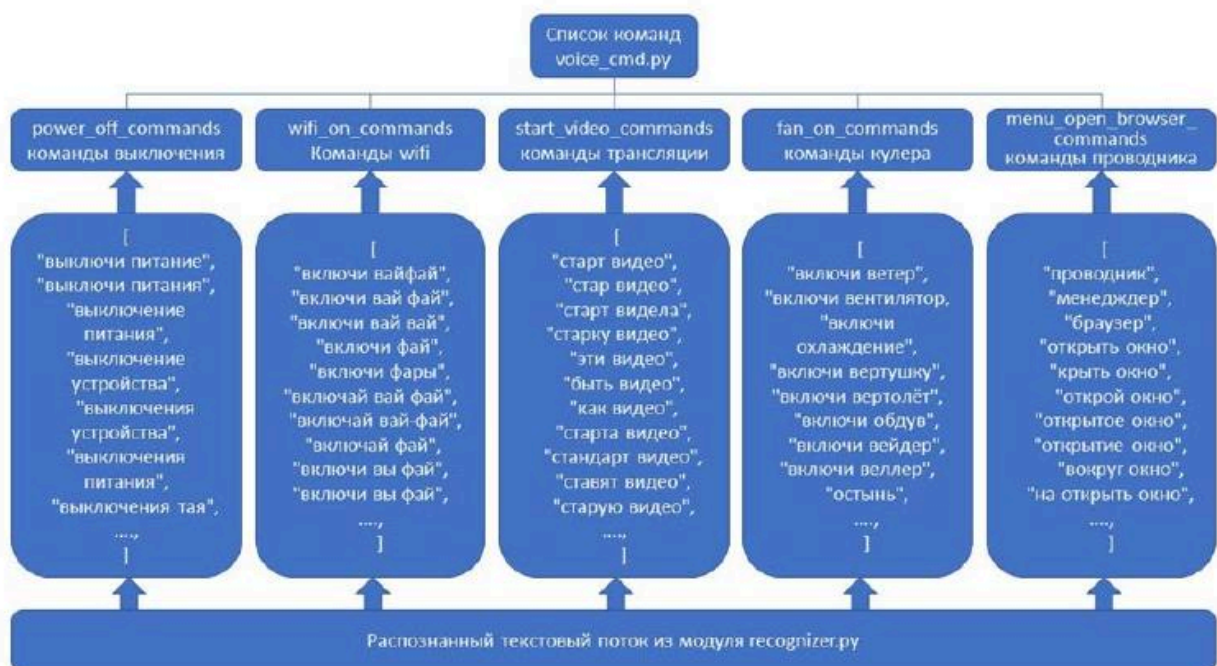


Рис. 3. Структура голосовых команд на примере некоторых из них

Проведена оценка точности для ряда команд, для которых собрана статистика правильного распознавания в условиях тишины, а также в условиях присутствия шума. Команды произносились на расстоянии 0,5 м от устройства. Параметры цифрового фильтра: порядок фильтра — 5, коэффициент усиления — 5. Производилось по 10 повторений каждой команды в условиях с шумом и без. Результаты показаны на рис. 3.

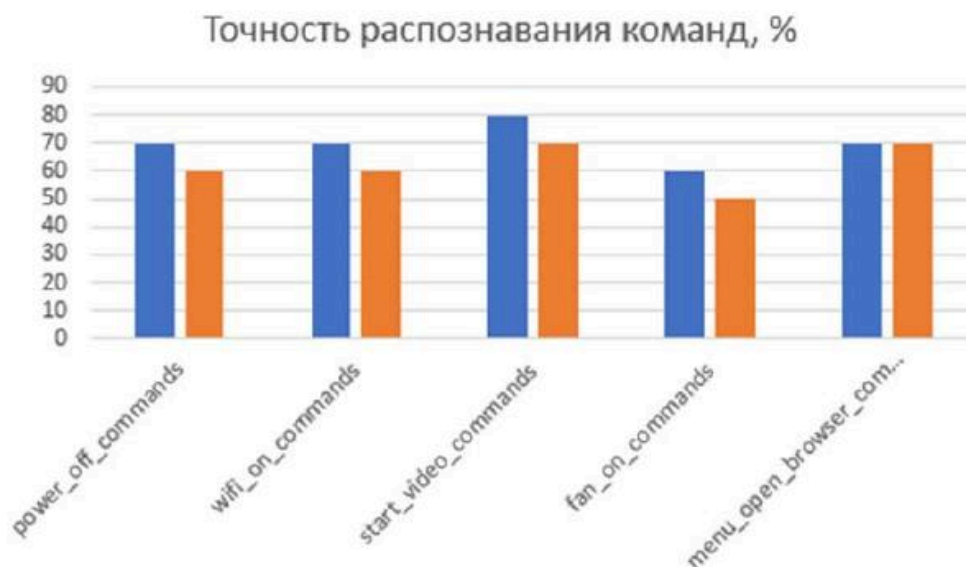


Рис. 4. Статистика точности распознавания отдельных команд в процентном соотношении, где синий — практически без фонового шума (30 дБ), оранжевый — с шумом (50 дБ). Громкость произношения команды около 60 дБ

Обсуждение полученных результатов. Результаты проверки на точность говорят о достаточно успешной реализации системы голосового управления на модуле-макете AR-гарнитуры. Значения точности распознавания сопоставимы с результатами, полученными на других платформах [6]. Обеспечение требуемой точности распознавания, зависит от таких параметров, как качество микрофонов и обработки звуковых сигналов, сложность используемой голосовой модели и количество вариантов произнесения конкретной команды. В перспективе для повышения качества голосового управления следует выбрать более качественные микрофоны и внедрить более сложные алгоритмы фильтрации. Также для увеличения числа распознаваний можно расширить список команд путем добавления команд с погрешностью распознавания, которые можно определить при изучении текстового потока с распознавателя.

Заключение. В рамках данной работы был успешно разработан макетный модуль для AR-гарнитуры на базе микроконтроллера BPI CM4, использующий голосовое управление при помощи библиотеки распознавания Vosk. Внедренный набор команд реализует весь требуемый функционал устройства. Таким образом, была продемонстрирована возможность реализации функций AR-гарнитуры на миниатюрном микропроцессорном модуле.

Список источников

- [1] Carmigniani J. et al. Augmented reality technologies, systems and applications. Multimedia tools and applications, 2011, vol. 51, pp. 341–377. <https://doi.org/10.1007/s11042-010-0660-6>
- [2] Nee A.Y.C. et al. Augmented reality applications in design and manufacturing. CIRP annals, 2012, vol. 61, no. 2, pp. 657–679.

-
- [3] Korkiakoski M., Alavesa P., Kostakos P. Preference in voice commands and gesture controls with hands-free augmented reality with novel users. *IEEE Pervasive Computing*, 2024, vol. 23, no. 1, pp. 1–26. <https://doi.org/10.1109/MPRV.2024.3364541>
 - [4] Zhao J. et al. Voice interaction for augmented reality navigation interfaces with natural language understanding. 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020, pp. 1–6. <https://doi.org/10.1109/IVCNZ51579.2020.9290643>
 - [5] Chan Z.Y., Shum P. Smart office: a voice-controlled workplace for everyone. *Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control (ISCSIC '18)*. New York, USA, 2018, pp. 1–5. <https://doi.org/10.1145/3284557.3284712>
 - [6] Ibrahim U.I., Ohize H., Umar U.A., Aliyu Y. Design and Construction of Voice-Controlled Home Automation using Arduino. *ABUAD Journal of Engineering Research and Development (AJERD)*. 2024, vol. 7, is. 1, pp. 144–152. <https://doi.org/10.53982/ajerd.2024.0701.14-j>