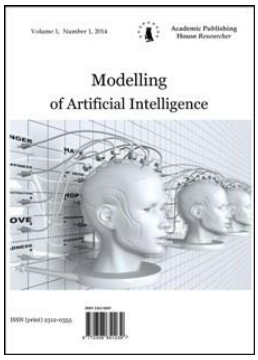

Published in the Russian Federation
Modeling of Artificial Intelligence
Has been issued since 2014.
ISSN: 2312-0355
E-ISSN: 2413-7200
Vol. 9, Is. 1, pp. 4-14, 2016

DOI: 10.13187/mai.2016.9.4
www.ejournal11.com

Articles and statements

UDC 004.934

Application of Multi- Tier Applications Technology Datasnap in Designing a System of Automatic Segmentation and Recognition of Sppech Signal

Yedilkhan N. Amirgaliyev [a, *], Timur R. Mussabayev [b], Rustam R. Mussabayev [a]

[a] Institute of Information and Computing Technologies, Almaty, Kazakhstan
[b] Kazakh National University named after al-Farabi, Almaty, Kazakhstan

Abstract

In this paper we will address current issues in the field of development and application of automatic identification systems and segmentation of speech signals. The basic criteria for the shortcomings of such systems were formulated. The review of the types of speech recognition systems was conducted, and the optimum architecture for them, including information used in leading IT companies was described. The possibility of using multi-tier architectures for solving problems of speech recognition and their advantages were considered. Also practical implementation of multi-tier architecture based on DataSnap technology in voice recognition system for geo search in Kazakh language was described.

Keywords: speech signal, speech recognition, speech segmentation, neural networks, Hadoop, MapReduce, multi-tier architecture, DataSnap.

1. Введение

Прежде чем начать детальное описание архитектуры системы, мы должны обратить внимание на некоторые проблемы и особенности, которые существуют в существующих системах автоматического распознавания речи (команд). Несмотря на то, что различными исследователями из разных стран проводились фундаментальные исследования по разработке и применению систем автоматического распознавания речи, до сих пор мы можем говорить о недостаточном прогрессе в этой области. В частности, это связано и с использованием общепринятых алгоритмов и невозможности выхода за их рамки. Некоторые исследования, проведенные в научном сообществе университета Карнеги Меллон показали, что универсальные системы распознавания речи, так и не преодолели уровень точности распознавания слитной речи больше чем в 80 %. Тогда как у человека этот показатель составляет 96–98 %. То есть точность систем распознавания речи осталась на уровне 1999 года, и с тех пор не улучшилась (Burger et al., 2006: 809-814). Отчасти, причиной такого явления может быть использование фон-Неймановской архитектуры в машинах, на

* Corresponding author
E-mail addresses: amir_ed@mail.ru (Ye.N. Amirgaliyev), tmusab@yandex.ru (T.R. Mussabayev), rmusab@gmail.com (R.R. Mussabayev)

которые ложится задача по распознаванию человеческой речи. Другая причина может быть в том, что речевой сигнал не несет достаточно информации для осуществления процесса преобразования речи в команды (текст). Отдельную проблему создает наличие шумов и искажений в речевом сигнале. Также влияют факторы, составляющие функциональность систем автоматического распознавания речевого сигнала, такие как распознавание в режиме реального времени, распознавание одновременно нескольких дикторов, и другие, которые могут накладывать ограничение на точность распознавания таких систем. На рис. 1 изображена классификация систем распознавания речи по нескольким критериям (Anusuya and Katti, 2009). Самая нижняя строчка критериев описывает систему с высокой точностью распознавания речи.

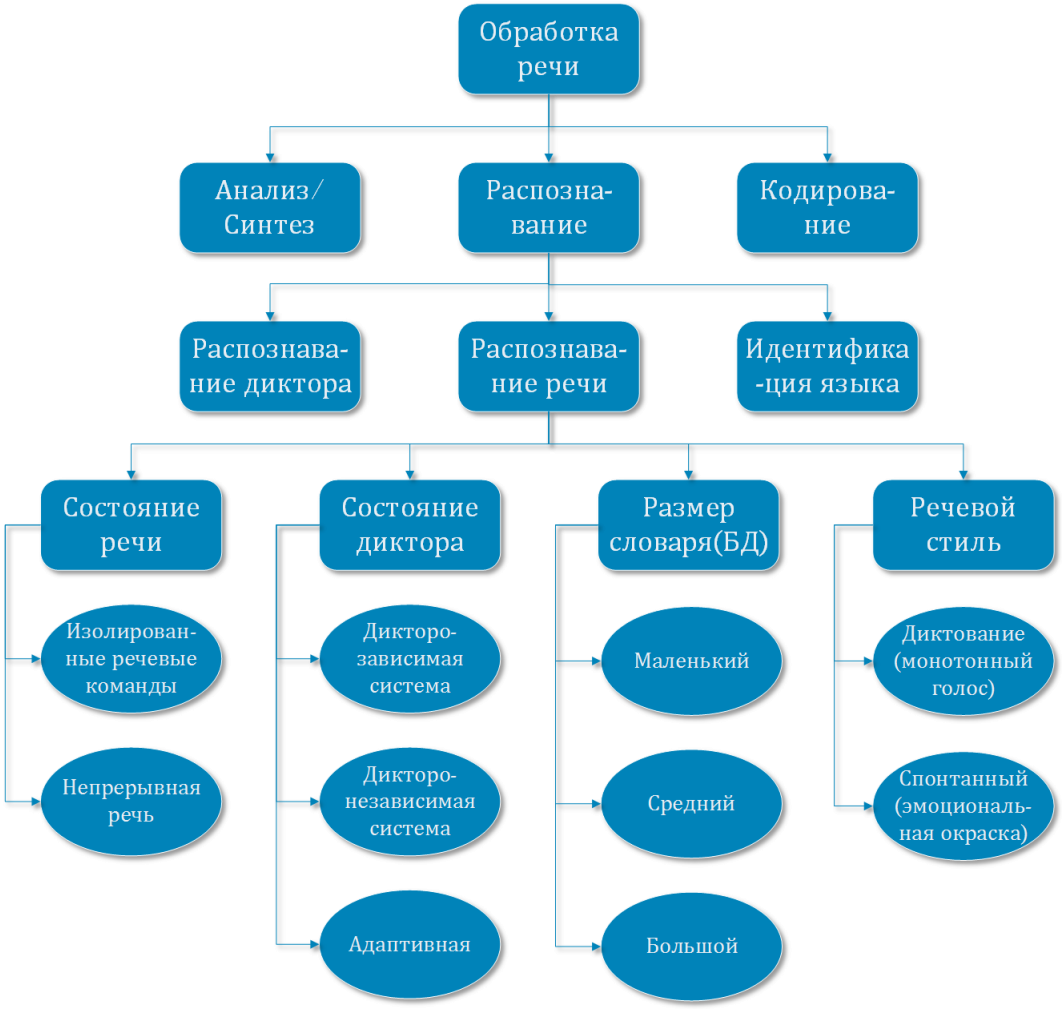

Рис. 1. Классификация систем распознавания речи

Однако если сравнивать человека и ЭВМ, то можно задать следующий вопрос. Каким же образом человеку удается воспринимать и анализировать человеческую речь с процентом точности недостижимым для машин? На данный момент, однозначного ответа на этот вопрос не получено. Но можно с уверенностью утверждать, что человек воспринимает речь учитывая другие факторы, сопутствующие моменту получения речевого сигнала, такие как текущую ситуацию, выражение лица и эмоции говорящего, тембр голоса и др. Несомненно также то, что человеческий мозг, по структуре очень похожий на 3D модель нейронной сети, позволяет фактически мгновенно получать визуальные ассоциации из произнесенной речи, а также дополнять не полную и даже с виду не связную речь. Другим не маловажным фактором является способность человека предсказывать воспринимаемое речевое сообщение и анализировать её семантическое содержание.

Все выше перечисленное приводит нас к нескольким взаимосвязанным между собой выводам, сформулированных на Рис. 2.

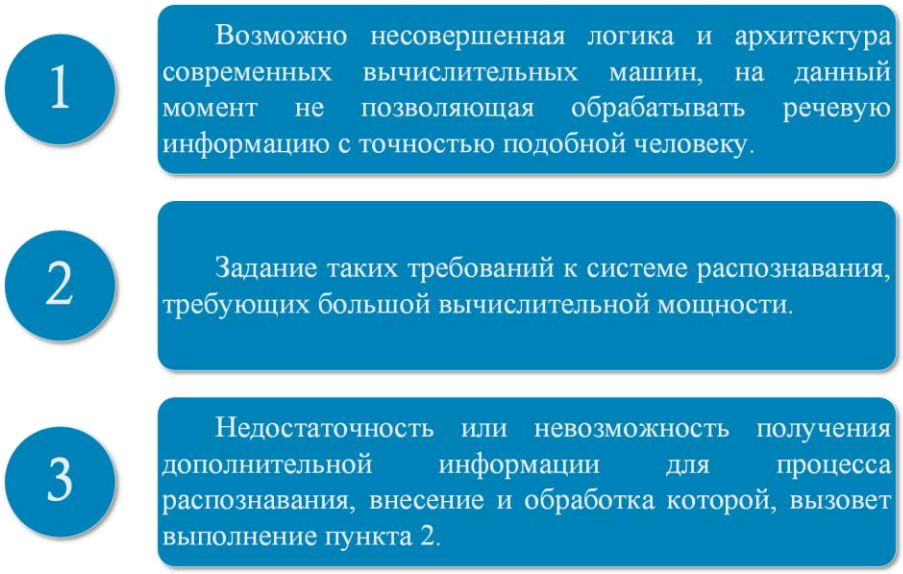

Рис. 2. Основные критерии недостатков современных систем распознавания речи

2. Результаты

2.1. Нейронная реализация. Решения описанных недостатков в полном объеме на сегодняшний день не существует. Однако есть варианты архитектуры, позволяющие организовать работу системы так, чтобы влияние вышеперечисленных недостатков было сведено к минимуму. Эффективным способом решения может быть построение модели восприятия информации, похожей на человеческую. Одной из таких первых и фундаментальных моделей восприятия информации мозгом (кибернетическая модель мозга) является персептрон. Персептрон предложил в 1957 году амер. ученый Ф. Розенблат. Персептрон стал одной из первых нейросетевых моделей. Его обучение заключается в последовательной коррекции положения разделяющей гиперплоскости по текущим результатам распознавания входных сигналов (Глушков и др., 1974: 156-158). Логическая схема трехслойного персептрона с двумя выходами изображена на Рис. 3.

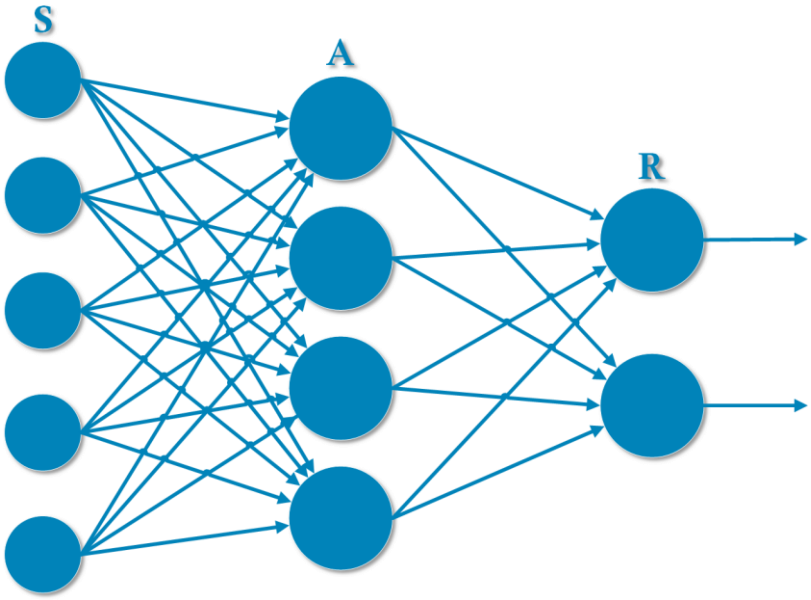

Рис. 3. Логическая схема персептрона

Однако применение нераспределенного персептрона на машинах фон-Неймана с принципами двоичного кодирования имеет некоторые недостатки. Даже при сравнительно малом числе элементов обучающей выборки для решения задач распознавания с помощью

трехслойного персептрона требуются физически нереализуемые объемы аппаратуры (по числу необходимых А-элементов и значениям весов связей) и длительности обучения (по числу отдельных коррекций весов). Были попытки улучшить рабочие характеристики персептрона путем усложнения его структуры (напр. переход к многослойным схемам) или путем усложнения процедуры обучения (напр., коррекцией весов других связей). При подобных усовершенствованиях теряются такие привлекательные стороны трехслойных персептронов, как простота и ясность схемной организации и процедуры обучения (Глушков и др., 1974).

2.2. Кластерная реализация. Возможным решением является ориентирование модели персептрона в основу сетевых методов планирования и управления с распределением не самой нейронной сети, а обучающих выборок. Массив обучающих векторов или входных данных разделяется на несколько блоков, а блоки в свою очередь распределяются между вычислительными нодами. Таким образом, все вычислительные действия распараллеливаются в рамках одной эпохи (Nguyen, 2013: 99-103). При этом, в процессе обработки входных данных используется простой подход, который называется «разделяй и властвуй». Решения с использованием данного подхода, и предложенное авторами, это использование сетевой многозвенной архитектуры.

Гиганты в области информационных технологий, такие как, Microsoft, Google и Apple уже используют подобные многозвенные архитектуры, в том числе и для задач распознавания. Частным случаем является использование распределённых облачных GRID систем с WEB-сервисами. Например, Apple Siri бекенд, запускает тысячи сервисов, и они работают на равном количестве вычислительных нодов. Вычислительные данные системы Siri находятся в HDFS (распределенной файловой системе Hadoop) кластера. Apple, создали собственную PaaS-подобную структуру планировщика под названием J.A.R.V.I.S., которая позволяет разработчикам внедрять услуги Siri очень доступным способом (What is Siris...). На Рис. 4 изображена обработка данных в кластере Hadoop по принципу «разделяй и властвуй» (Big data...). Очевидно сходство архитектур персептрона и кластера Hadoop.

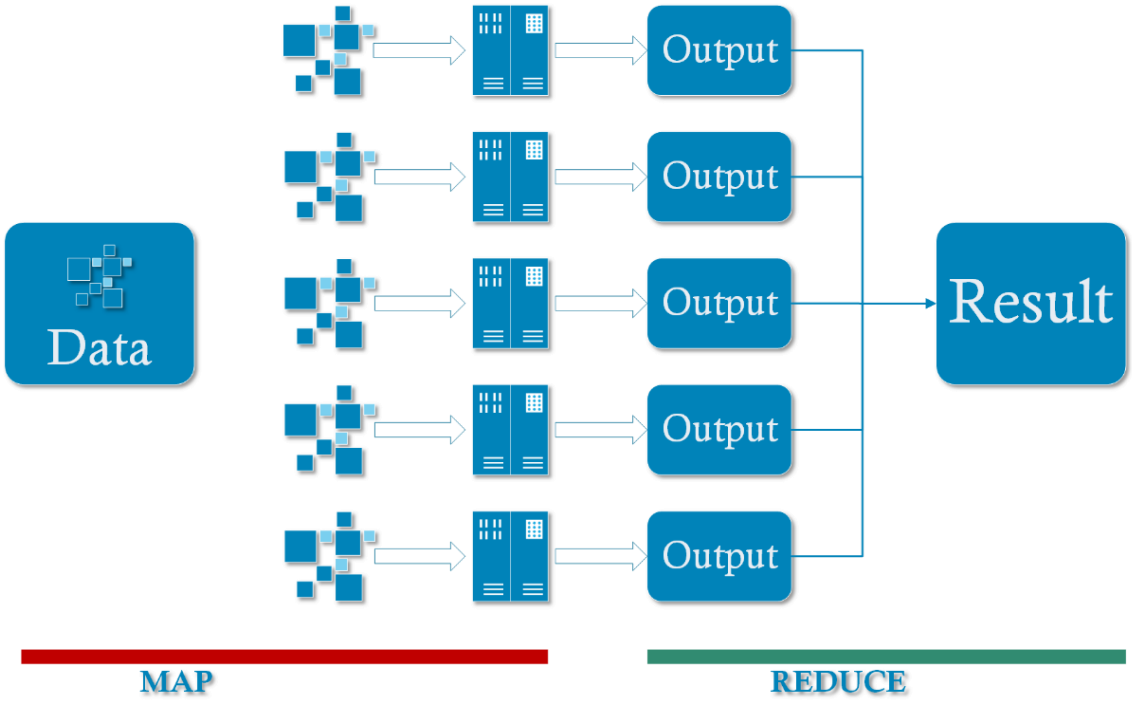

Рис. 4. Обработка данных в кластере Hadoop по принципу «разделяй и властвуй»

2.3. Многозвенная архитектура. Надо сказать, что область обработки речевого сигнала являются очень подходящими для применения в многозвенных архитектурах. Так как задачи подобного рода легко разбиваются на несколько независимых подзадач. Автором

уже были описан алгоритм автоматической сегментации речевого сигнала для задачи распознавания речевого сигнала в текст (Амиргалиев и др., 2015: 37-44). Основные этапы изображены на рис. 5.

Рис. 5. Алгоритм автоматической сегментации речевого сигнала

Построение эффективной сетевой многозвенной архитектуры будет зависеть от типа распознающей речевой системы и от объема речевых данных. Однако, в любом из случаев, будут присутствовать звенья пред- и постобработки, а также «круг звеньев» по числу независимых вычислительных задач. В нашем случае, под «кругом звеньев» подразумевается объединение независимых вычислительных машин по сетевой топологии позволяющей одновременно получать и независимо обрабатывать речевой сигнал со звена А (предобработка) к направлению звена В (постобработка). Необязательно для каждой задачи выделять ресурсы независимой машины. Каждая из машин, если она представлена в виде мультикомпьютера, может выполнять одновременно несколько задач посредством потоковой обработки. Самым «тяжелым» этапом с точки зрения вычислений является параметризация речевого сигнала, которая и будет выполняться в «круге звеньев». Подобная архитектура представлена на Рис. 6.

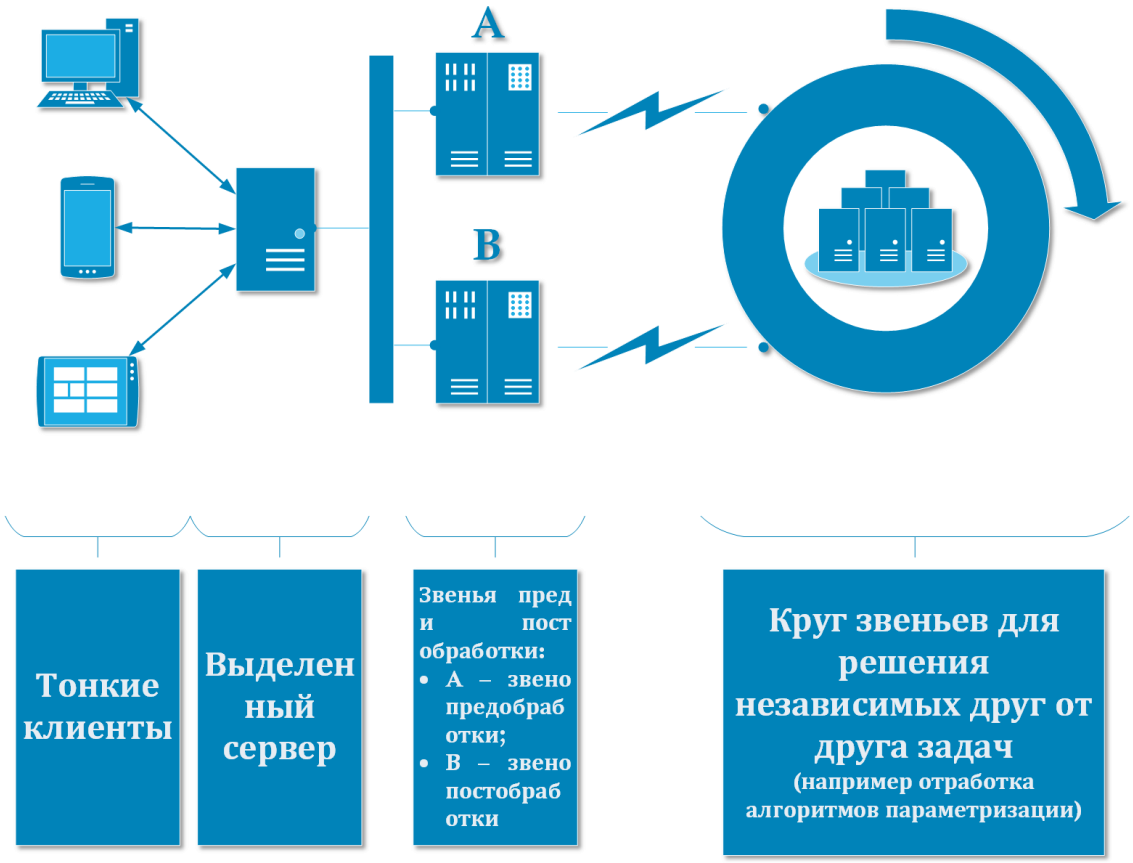

Рис. 6. Сетевая схема взаимодействия в многозвенной архитектуре

В многозвенных архитектурах, в рамках оптимизации под аппаратное обеспечение, позволено исключать, увеличивать или объединять звенья. Однако минимальное количество звеньев должно быть не менее трех. Реализовать подобную архитектуру можно посредством простой и популярной технологии распределенных многозвенных приложений DataSnap. Данная технология кроссплатформенна, имеет механизмы обратного вызова, поддерживает возможности изменения серверных частей без необходимости перетрансляции клиентского приложения, поддерживает сетевые протоколы безопасности, а также механизмы аутентификации и авторизации (Баженова, 2003; Осипов, 2012).

Авторами была разработана система распознавания изолированных речевых команд казахского языка с разделением программной функциональности по трехзвенной архитектуре на базе технологии DataSnap.

2.4. Описание сервера приложений. Ниже, на Рис. 7 представлен скриншот разработанного серверного приложения системы распознавания речевых команд казахского языка. Оно состоит из двух графических окон. Слева это интерфейсный сервер обмена информацией DataSnap. В данном приложении обмен информацией осуществляется через сетевой порт 8086. Процесс генерации и обработки данных будет рассмотрен далее. Сервер приложений DataSnap активируется автоматически с запуском основной программы.

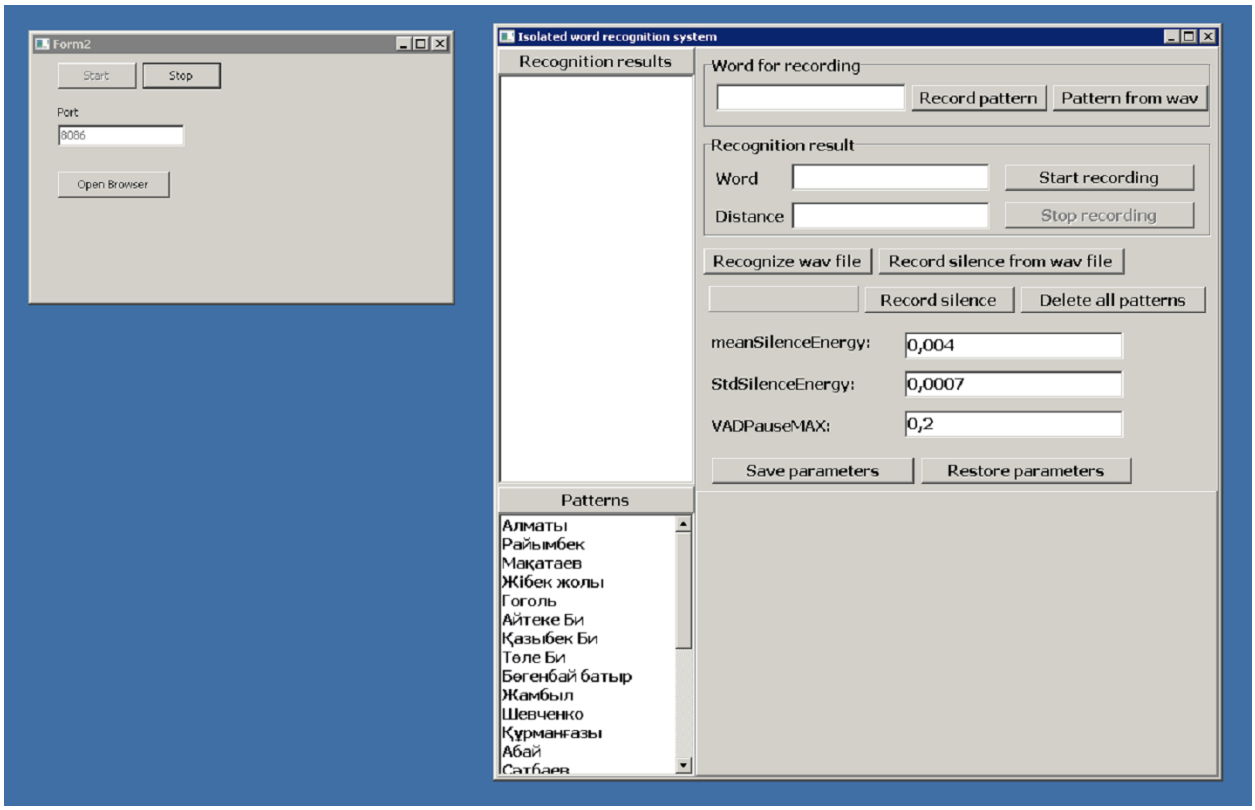

Рис. 7. Вид сервера-приложений системы распознавания речевых команд казахского языка

В главном окне справа представлен главный графический интерфейс этой системы распознавания речевых команд казахского языка. Он состоит из нескольких частей. В окне можно увидеть два поля: “Recogntion results” и “Patterns”. Первый это текстовое поле в виде списка используется для внесения в него уже распознанных системой речевых слов соответствующих речевым паттернам из второго поля. Поле «Patterns» содержит в себе уже надиктованные речевые паттерны, хранящиеся в файле Patterns.dat, который находится рядом с исполняемым файлом программы и используется как эталон для сравнения с уже поступающими на обработку речевыми сигналами.

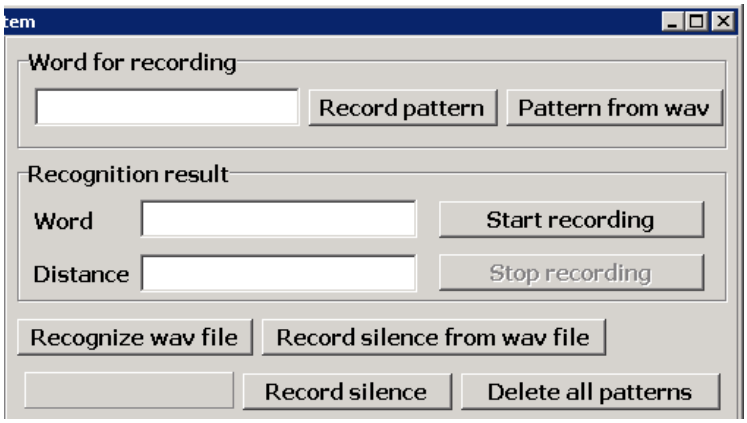

Рис. 8. Управляющая панель системы распознавания речевых команд

В верхней части программы, которая изображена на Рис. 8, производится управление речевыми паттернами на сервере приложений. Здесь можно записывать и удалять новые речевые паттерны, параметры тишины в помещении и др.

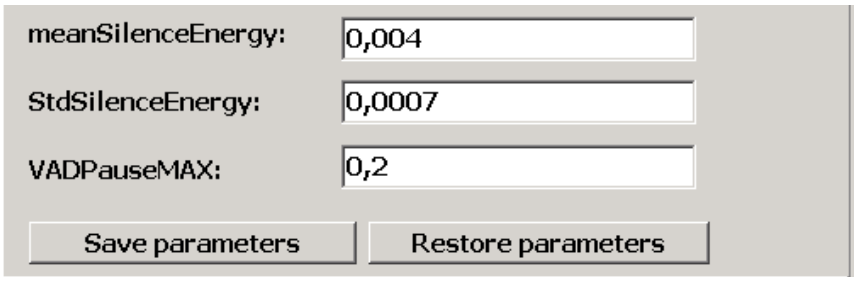

Рис. 9. Параметры системы распознавания речевых команд

Другой важной частью системы является панель задания параметров распознавания речевых команд (изображена на Рис. 9). Здесь задаются такие параметры как размер речевой паузы между произнесёнными словами, значения энергии тишины и др.

Возможно использование дополнительных звеньев для улучшения характеристики системы распознавания речевых команд, такие как лингвистические процессоры, а также отдельное звено для хранения и обработки речевых паттернов.

2.5. Описание тонкого клиента. Теперь рассмотрим работу тонкого клиента (Рис. 10) системы распознавания речевых команд написанного для планшета на базе операционной системы Android - 4.4 KitKat. Речевыми командами осуществляется управление модулем карт «Google Map», который позиционирует искомое местоположение при помощи голосовых команд казахского языка и выводит детальную карту искомого объекта. Здесь присутствует 4 режима работы приложения. Начать запись для распознавания, остановить запись, прослушать, остановить запись и отправить на сервер приложений для обработки. Также интерфейс системы поддерживает три языка: английский, казахский и русский.

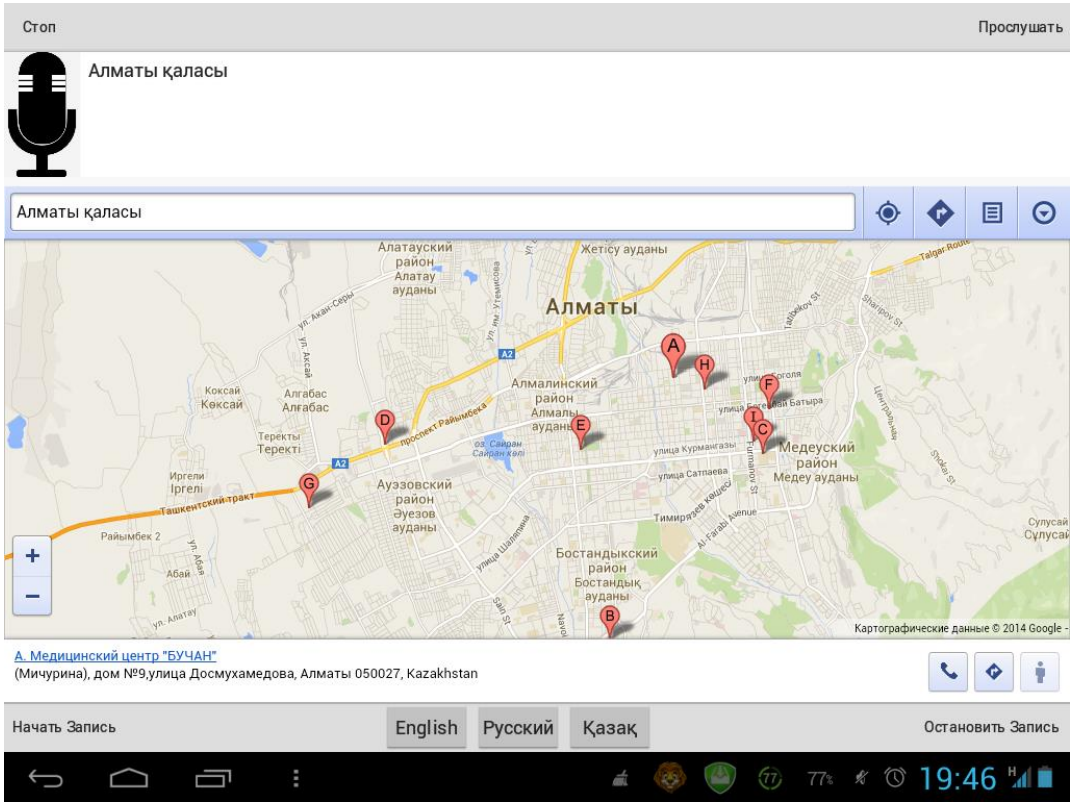

Рис. 10. Тонкий мобильный клиент системы распознавания речевых команд на казахском языке

Тонкий клиент не осуществляет вычислительных операций по распознаванию казахской речи. При инициации записи звука «KazSpeech» обращается к микрофону на мобильном устройстве и записывает звук с микрофона в звуковой файл, который в свою очередь отсылается на сервер приложений с помощью технологии DataSnap. Сервер получает файловые данные и записывает их в предварительно созданный каталог, имя каталога составляет не менее 15 случайно подобранных букв и цифр разного регистра. Если нужно, то полученный звуковой файл декодируется сервером приложений в удобное для обработки представление и осуществляет распознавание произнесенной казахской речи по описанным выше алгоритмам. Полученный текстовый результат сервер приложений отправляет обратно клиентской программе и уничтожает ранее полученный от клиента звуковые данные для освобождения необходимого вычислительного пространства. Клиентская мобильная программа при получении распознанного казахского текста инициирует поиск в картах Google и позиционирует ранее произнесенное вслух необходимое местоположение. Это наглядный пример того, как посредством произнесения речевых команд можно эффективно управлять какой-либо программой или устройством.

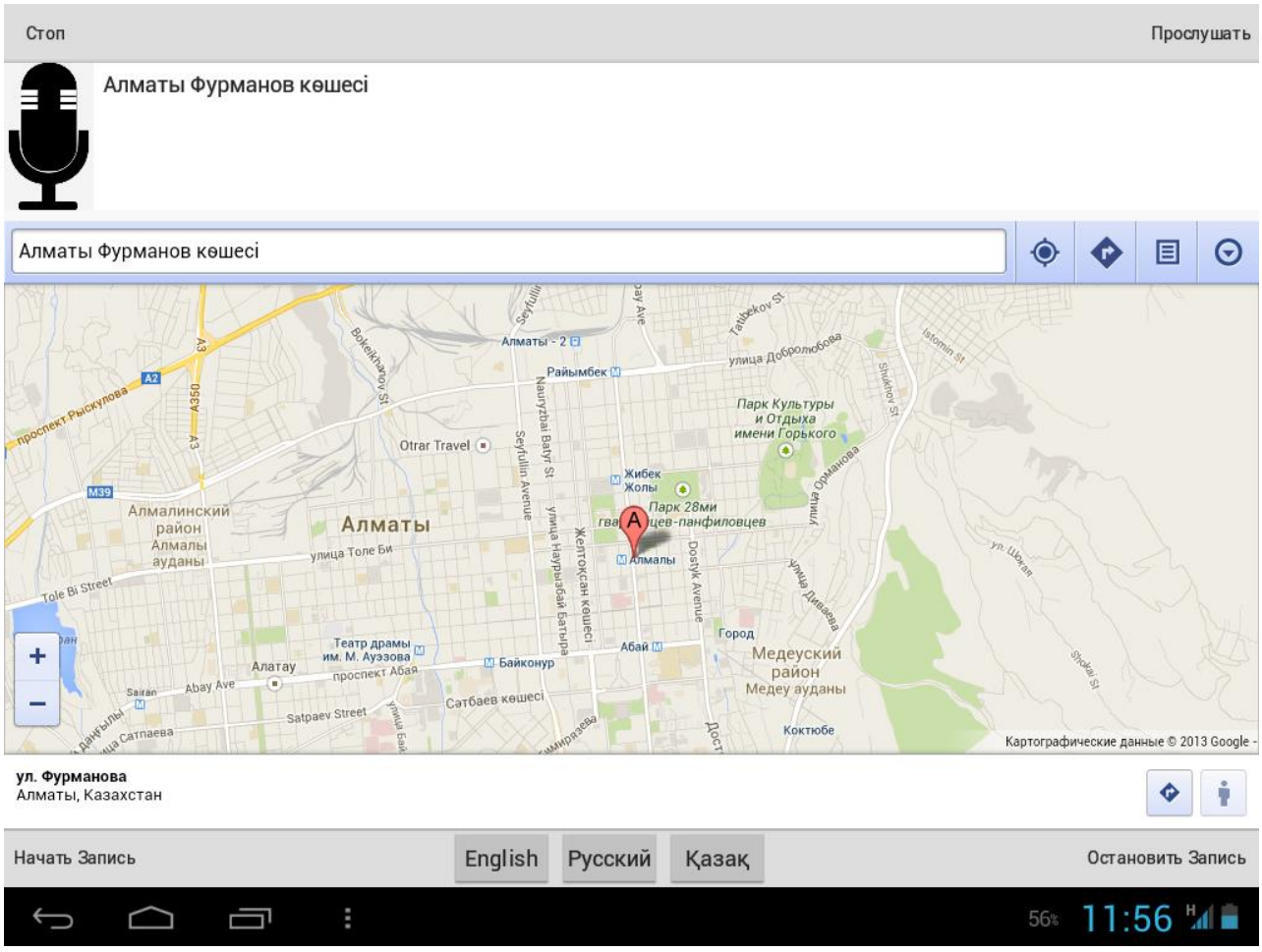

Рис. 11. Тонкий мобильный клиент системы распознавания речевых команд на казахском языке

3. Заключение

Согласно результатам научных исследований, в настоящее время, универсальные системы распознавания речи не обладают достаточной высокой точностью распознавания на уровне человеческого. Автором были сформулированы основные критерии существующих недостатков систем распознавания речи. Предложены оптимальные архитектуры построения таких систем, для преодоления обозначенных недостатков, в том числе используемые в настоящее время в ведущих IT-компаниях. Описаны возможности

использования многозвенных архитектур в системах распознавания и сегментации речи. Практически реализована и рассмотрена система распознавания речевых команд казахского языка на основе технологии DataSnap.

Литература:

Амиргалиев и др., 2015 - Амиргалиев Е.Н., Мусабаев Р.Р., Мусабаев Т.Р. Автоматическая сегментация речевого сигнала на окна со стабильными спектральными характеристиками на основе кратковременных алгоритмов анализа синхронизированных с частотой основного тона // Труды XI Международной Азиатской школы-семинара «Проблемы оптимизации сложных систем», Часть I, Кыргызская республика, г. Чолпон-Ата, 2015. С. 37–44.

Баженова, 2003 - Баженова И.Ю. Delphi 7 Самоучитель программиста. М.: Кудиц-Образ, 2003. 349 с.

Глушков и др., 1974 - Глушков В.М., Амосов Н.М., Артеменко И.А. Энциклопедия кибернетики. Том 2. К.: Главная редакция украинской советской энциклопедии, 1974. С. 156–158.

Осипов, 2012 - Осипов Д. Delphi XE2. Санкт-Петербург: БХВ-Петербург, 2012. 912 с.

Anusuya and Katti, 2009 - Anusuya M.A., Katti S.K. Speech Recognition by Machine: A Review // (IJCSIS) International Journal of Computer Science 6, No. 3, 2009. pp. 181–205.

Big data… - Big data what is hadoop // http://singhvikash.blogspot.co.uk/2013/12/big-data-what-is-hadoop.html

Burger et al., 2006 - Susanne Burger, Zachary A. Sloane and Jie Yang. Competitive Evaluation of Commercially Available Speech Recognizers in Multiple Languages // Italy, Genoa, Proceedings of LREC2006, 2006. pp. 809–814.

Nguyen, 2013 - Nguyen Zhang. A distributed platform for parallel training of disann artificial neural networks// International Journal of software products and systems, 2013. Iss. 3, pp. 99–103.

What is Siris… - What is Siris architecture in Apples data center // https://www.quora.com /What-is-Siris-architecture-in-Apples-data-center

References:

Amirgaliev i dr., 2015 - Amirgaliev E.N., Musabaev R.R., Musabaev T.R. Avtomaticheskaya segmentatsiya rechevogo signala na okna so stabil'nymi spektral'nymi kharakteristikami na osnove kratkovremennykh algoritmov analiza sinkhronizirovannykh s chastotoi osnovnogo tona [Automatic segmentation of a speech signal with stable window based on the spectral characteristics of transient analysis algorithms synchronized with the frequency of the fundamental tone] // Trudy XI Mezhdunarodnoi Aziatskoi shkoly-seminara «Problemy optimizatsii slozhnykh sistem», Chast' I, Kyrgyzskaya respublika, g. Cholpon-Ata, 2015. S. 37–44.

Bazhenova, 2003 - Bazhenova I.Yu. Delphi 7 Samouchitel' programmista [Delphi 7 Tutorial of the programmer]. M.: Kudits-Obraz, 2003. 349 s.

Glushkov i dr., 1974 - Glushkov V.M., Amosov N.M., Artemenko I.A. Entsiklopediya kibernetiki [Encyclopedia of Cybernetics]. Tom 2. K.: Glavnaya redaktsiya ukrainskoi sovetskoi entsiklopedii, 1974. S. 156–158.

Osipov, 2012 - Osipov D. Delphi XE2. Sankt-Peterburg: BKhV-Peterburg, 2012. 912 s.

Anusuya and Katti, 2009 - Anusuya M.A., Katti S.K. Speech Recognition by Machine: A Review // (IJCSIS) International Journal of Computer Science 6, No. 3, 2009. pp. 181–205.

Big data… - Big data what is hadoop // http://singhvikash.blogspot.co.uk/2013/12/big-data-what-is-hadoop.html

Burger et al., 2006 - Susanne Burger, Zachary A. Sloane and Jie Yang. Competitive Evaluation of Commercially Available Speech Recognizers in Multiple Languages // Italy, Genoa, Proceedings of LREC2006, 2006. pp. 809–814.

Nguyen, 2013 - Nguyen Zhang. A distributed platform for parallel training of disann artificial neural networks// International Journal of software products and systems, 2013. Iss. 3, pp. 99–103.

What is Siris… - What is Siris architecture in Apples data center // https://www.quora.com /What-is-Siris-architecture-in-Apples-data-center

УДК 004.934

Применение технологии многозвенных приложений DATASNAP при разработке системы автоматической сегментации и распознавании речевого сигнала

Едильхан Несипханович Амиргалиев [a, *], Тимур Рафикович Мусабаев [b],
Рустам Рафикович Мусабаев [a]

[a] Институт информационных и вычислительных
технологий КН МОН РК, Алматы, Казахстан
[b] Казахский национальный университет им. ал-Фараби, Алматы, Казахстан

Аннотация. В данной работе рассматриваются существующие проблемы в области разработки и применения систем автоматического распознавания и сегментации речевых сигналов. Сформулированы основные критерии для оценки недостатков данных систем. Проведен обзор типов речевых систем распознавания, а также описаны их оптимальные архитектуры для них, в том числе используемые в ведущих ИТ-компаниях. Рассмотрена возможность использования многозвенных архитектур для решения задач распознавания речи и их преимущества с описанием практической реализации на основе технологии DataSnap на примере системы распознавания речевых команд для геопоиска на казахском языке.

Ключевые слова: речевой сигнал, распознавание речи, сегментация речи, нейронные сети, Hadoop, MapReduce, многозвенная архитектура, DataSnap.

[*] Корреспондирующий автор
Адреса электронной почты: amir_ed@mail.ru (Е.Н. Амиргалиев),
tmusab@yandex.ru (Т.Р. Мусабаев), rmusab@gmail.com (Р.Р. Мусабаев)