ПРИМЕНЕНИЕ ИМИТАЦИОННОГО МОДЕЛИРОВАНИЯ ДЛЯ ОЦЕНКИ КАЧЕСТВА ПРИМЕНЯЕМЫХ МЕТОДОВ РАСПОЗНАНИЯ РЕЧИ В АВТОМАТИЗИРОВАННОЙ СИСТЕМЕ ВЕДЕНИЯ ДИАЛОГОВ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

А.А. Тарасьев, М.Е. Филиппова, К.А. Аксенов, Е.Н. Таланцев, И.А. Калинин (Екатеринбург)

Автоматическое распознавание речи является одной из ключевых задач при построении систем взаимодействия человека и машины на основе речевого интерфейса. При этом на сегодняшний день наиболее востребованными и вместе с тем наиболее сложными в реализации остаются системы распознавания спонтанной речи. Сложность построения таких систем вызвана такими особенностями, как значительная изменчивость темпа речии психофизическое состояние диктора (манера произнесения фраз, эмоциональное состояние, кашель, заикания), наличие акцентов или большое количество используемых словоформ.

Задача дополнительно осложняется наличием пауз, повторов, не лексическими вставками, словами-паразитами и прочим. На сегодняшний день разработано большое количество методов распознавания речи с учетом описанных ограничений спонтанной речи, а также существует большое количество речевых движков с открытым исходным кодом и коммерческих, которые могут служить основой для подобных систем.

Однако существующие системы распознавания речи обладают недостатками при распознавании как речи в целом, так и отдельных языковых единиц.

Таким образом, при построении модуля распознавания речи разрабатываемой голосовой системы ведения диалогов в режиме реального времени «ТВИН»решения принимались на основании идеи использования, адаптации и доработки существующих и хорошо себя зарекомендовавших подходов, а не создания концептуально новых алгоритмов.

На сегодняшний день системы Google и Yandex демонстрируют самую высокую точность распознавания слитной русской речи - около 85% [8-9]. Такое качество распознавания обеспечивается, во-первых, огромными наборами акустических данных для обучения (тысячи часов речи), а, во-вторых, присутствием многих запрашиваемых фраз и словоформ из текстовых поисковых запросов, на которых обучались языковые модели. За счет интеграции данных двух систем была реализована собственная система распознавания речи.

На базе существующих систем распознавания речи для системы был подобран и реализован следующий вариант речевого движка. Архитектура разработанной системы распознавания представляет собой мультиагентную систему (МАС) использующую для задачи распознавания комбинацию двух систем GoogleSpeechRecognition и YandexSpeechKitCloud и является оптимальным решением для голосовой вопросно-ответной системы.

Модуль распознавания системы «ТВИН» состоит из трех основных подсистем:

1. Виртуальная АТС – Реализует функционал принятия звонка и маршрутизацию трафика до подсистемы распознавания речи;
2. Подсистема Распознавания речи – программный комплекс, основной задачей которого является переадресация трафика до требуемой системы распознавания;
3. Модуль принятия решения – программный комплекс, состоящий из авторских алгоритмов обработки текстовой информации. Оснащен подпрограммой принятия решения и подпрограммой синтеза речи.

Выбор системы распознавания может быть настроен заранее, либо определяться динамически с использованием модуля поддержки принятия решений.

Решение использовать обе популярные системы распознавания речи вызвано несколькими факторами:

1. Данные системы являются закрытыми, вследствие чего невозможно однозначно полагаться на качество распознавания каждой.
2. Качество распознавания отдельных языковых структур варьируется для данных систем.

3. Данные системы используют различные внутренние механизмы распознавания, вследствие чего могут по-разному генерировать конечный результат, что может быть использовано для различных предметных областей (в случае с явной настройкой системы распознавания на этапе проектирования сценариев диалога).
4. Данные системы предлагают различающийся дополнительный функционал, который также может быть вариативно применен для разных областей использования.
5. Данные системы различны по стоимости использования, что позволяет в некоторых предметных областях использовать более дешевые решения с более простой инфраструктурой.

Из перечисленных пунктов наиболее спорным и требующим внимания является утверждение о различном качестве распознавания различных языковых структур. Вследствие этого необходимо произвести тестирование качества распознавания некоторых базовых структур речи для обеих систем.

Для данного тестирования не может быть использован традиционной подход анализа качества распознавания изолированных дикторонезависимых фраз, поскольку подобная аналитика не может быть репрезентативной. Это связано в первую очередь с особенностями языковых моделей, наличием паронимов, вариативном произношении слов в различных ситуациях или различными людьми, наличием шумов, длинной, сложностью фраз, наличием эмоциональной окраски и т.п.

Таким образом, для решения проблемы необходимо применение комплексного подхода, основанного на проведении большого числа экспериментов с использованием имитационного моделирования.

Моделироваться при этом будут реальные сценарии диалогов в течение времени.

Постановка эксперимента

На основании вышесказанного, для тестов можно выделить следующие критерии качества распознавания речи:

1) Процент распознанных коротких эмоциональных фраз.
2) Процент распознанных длинных фраз.
3) Процент распознанных терминов, характерных для предметной области.
4) Процент распознанных имен собственных.
5) Процент распознанных простых числительных.
6) Процент распознанных сложных числительных.
7) Процент распознанных дат, адресов и прочей аудиоинформации, содержащей числительные.
8) Процент распознанной речи в условиях шума и других искажений.

Проблемным вопросом в контексте данной задачи является способ построения моделей, основанных на реализации конкретных сценариев диалога. Традиционные системы моделирования не имеют в своей реализации подобного функционала, ввиду узкой направленности данной проблемы.

Методика эксперимента

Система «ТВИН» имеет в своей реализации модуль визуальной настройки сценариев диалогов – скриптов (Рисунок 1) [2-4, 6].

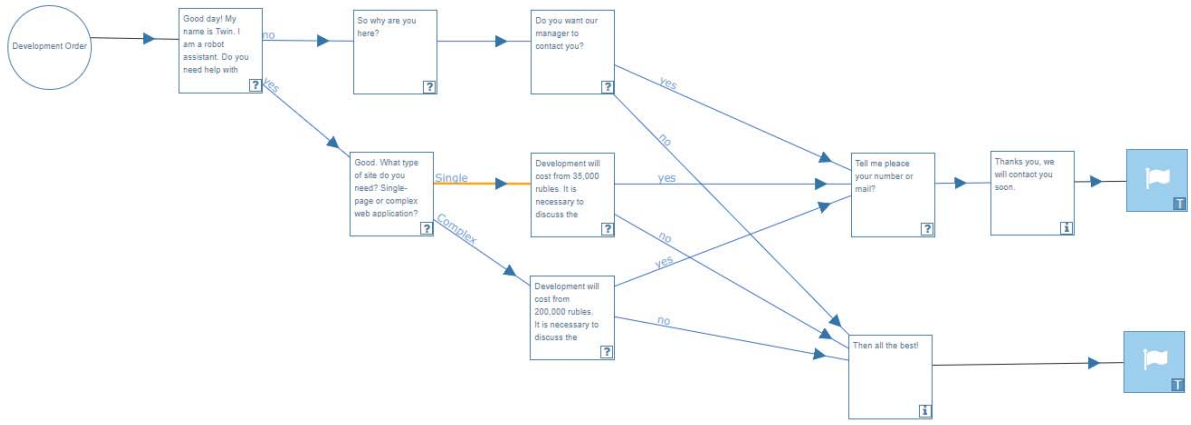

Рисунок 9– Пример скрипта сценария диалога в системе «ТВИН»

Для решения проблемы организации имитационных экспериментов был разработан специализированный комплекс для генерации (воспроизведения диалогов) на основании уже существующих в системе технологий.

Идея заключалась в организации автоматических диалогов по заранее описанным возможным сценариям между роботами. То есть необходимо составить несколько пар сценариев для исследования каждого выделенного критерия.

Базируясь на данных технологиях было разработано несколько скриптов для моделирования и тестирования выделенных лексических единиц. Особое внимание при этом уделялось качеству распознавания числительных, дат и адресов.

К данному модулю были подключены разные рассматриваемые вопросно-ответные системы распознавания речи.

Процесс моделирования заключался в многократном автоматическом прогоне заранее подготовленных аудиофайлов для первого скрипта, которые были предварительно протегированы экспертами по выделенным лексическим категориям. При этом изменялись настройки распознавания. По мере продвижения диалога по сценарию на основании корректности распознавания вторым участником беседы диалог уходил в нужную прогнозируемую ветку сценария или нет. На основании статистических данных по окончанию моделирования записывались результаты.

Обсуждение результатов

Результаты тестирования представлены на следующих диаграммах (Рисунок 2-9) [1,7].

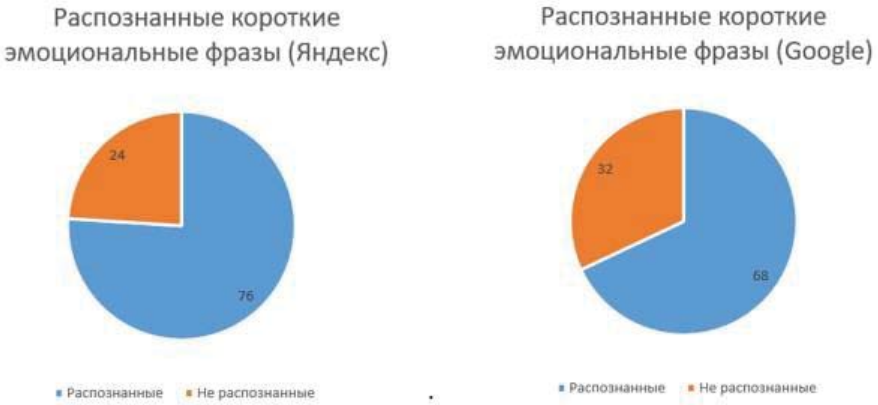

Рисунок 10 – Распознанные короткие эмоциональные фразы

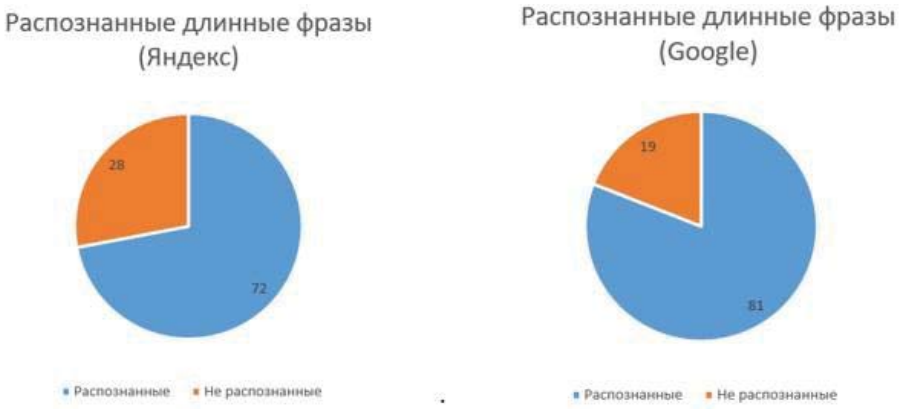

Рисунок 11 – Распознанные длинные фразы

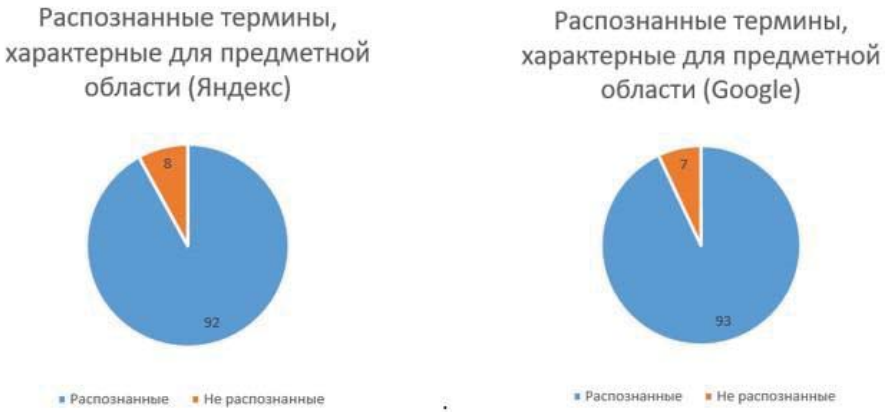

Рисунок 12 – Распознанные термины

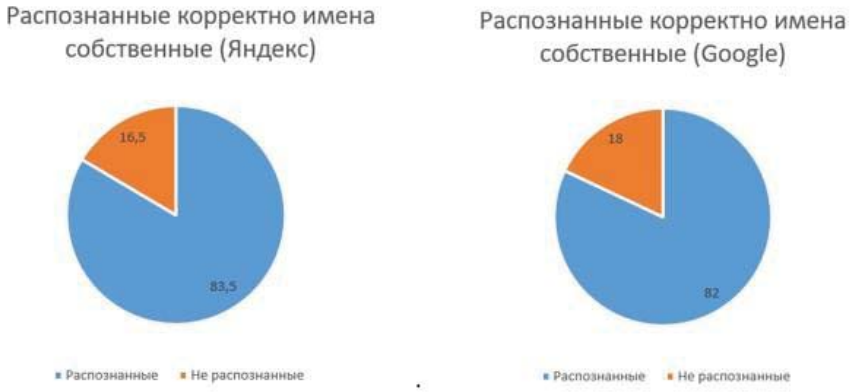

Рисунок 13 – Распознанные имена собственные

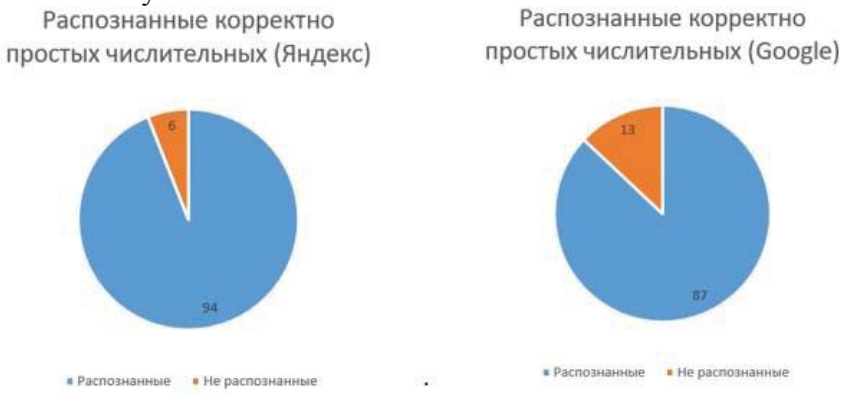

Рисунок 14 – Распознанные простые числительные

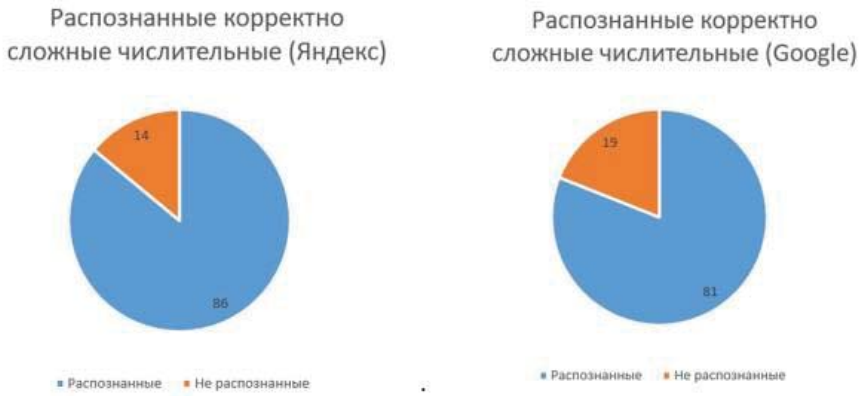

Рисунок 15 – Распознанные сложные числительные

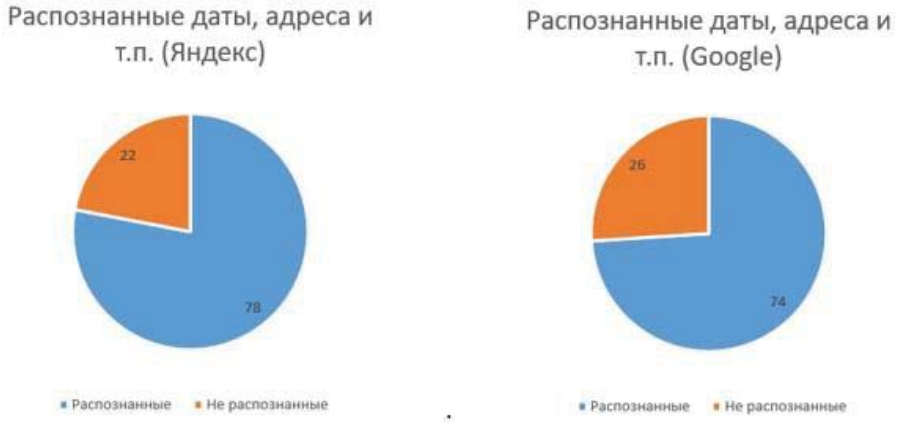

Рисунок16 – Распознанные даты, адреса и т.п.

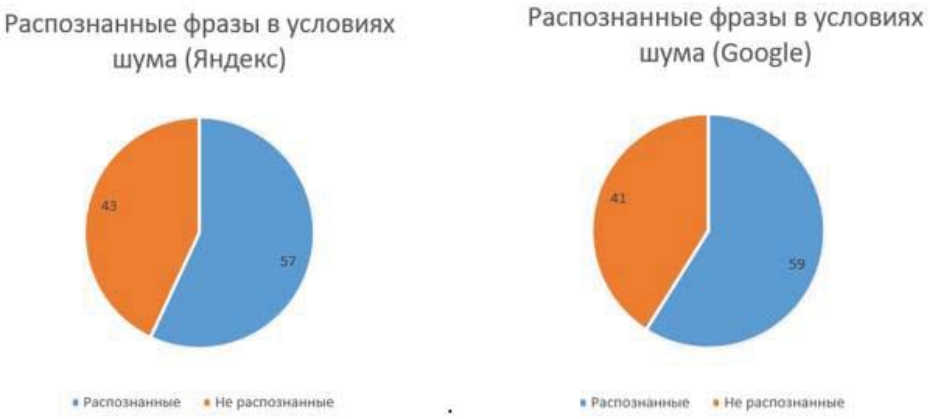

Рис. 17 – Распознавание в условиях шума

На основании полученных сведений можно сделать вывод о том, что система компании Яндекс лучше распознает короткие экспрессивные фразы, а также числительные, в то время как API Google лучше распознает длинные фразы и термины.

При этом обе системы имеют проблемы с распознаванием в условиях шума, неспокойного темпа речи и голосовых дефектах собеседника.

Это связано с тем, что обе системы лучше распознают фонемы и фразы, и хуже отдельные звуки, особенно в случае наличия шумов и других факторов, искажающих качество передаваемого аудиосообщения. Данное замечание подтверждается выводами, полученными другими независимыми исследователями [5].

На основании полученных данных был сформирован алгоритм работы системы распознавания речи, входящей в комплекс «ТВИН». Также в системы включен модуль последующей обработки распознанного текста – нормализация, выделение ключевых слов и т.п.

Применение данного модуля значительно упрощает конечное восприятие роботом произнесенной собеседником фразы и выбор последующих действий (произнесения соответствующих реплик), предусмотренных заданным сценарием[7].

Предложенный метод был реализован в модулях предварительной и последующей стороннему распознанию обработки. Полученный модуль распознавания речи также был протестирован с помощью используемых ранее моделей.

Качество распознавания в этом случае возросло по тестируемым показателям. В таблице 1 приведена статистика распознавания фраз системой «ТВИН» фраз по категориям.

Таблица 1 – Статистика распознавания фраз системой «ТВИН» по категориям

Тип фразы	Кол-во тестируемых фраз	Кол-во корректно распознанных фраз	Кол-во ошибок распознавания звуков, слогов	Кол-во ошибок распознавания форм слов	Кол-во ошибок распознавания полностью слов	Процент ошибок распознавания
Короткие эмоциональные фразы	4200	3402	187	529	82	19
Длинные фразы	3699	3107	104	471	17	16,00432549
Термины, характерные для предметной области	1216	1143	31	35	7	6,003289474
Имена собственные	2540	2159	249	46	86	15
Простые числительные	5366	5205	25	62	74	3,000372717
Сложные числительные	2200	1958	95	59	88	11
Даты, адреса и прочая аудиоинформация, содержащая числительные	1289	1006	163	86	34	18,00254453
Фразы в условиях шума и искажений	1659	553	436	391	279	40

Качество распознавания в этом случае возросло относительно лучших показателей сторонних сервисов по следующим показателям:

Процент распознанных коротких эмоциональных фраз – 5%.

Процент распознанных длинных фраз – 3%.

Процент распознанных терминов, характерных для предметной области – 1%.

Процент распознанных имен собственных – 1.5%.

Процент распознанных простых числительных – 3%.

Процент распознанных сложных числительных – 2%.

Процент распознанных дат, адресов и прочей аудиоинформации, содержащей числительные – 4%.

Процент распознанных фраз в условиях шума – 1%.

Заключение

Модуль распознавания речи в системе «ТВИН» использует в своей работе интеграцию двумя наиболее проработанными в настоящий момент существующими решениями YandexSpeechKit и GoogleSpeech API.

На основе применения имитационного моделирования было проведено тестирование используемых систем распознавания речи модуля распознавания речи "ТВИН". Для этого был реализован дополнительный модуль воспроизведения диалогов.

На основании полученных сведений можно сделать вывод о том, что система компании Яндекс лучше распознает короткие экспрессивные фразы, а также числительные. Напротив, API Google лучше распознает длинные фразы и термины.

На основании полученных сведений были созданы модули предварительной обработки, динамического выбора системы распознавания и последуюзей обработки распознанного текста. Качество распознавания интегрального решения – модуля распознавания речи системы "ТВИН" существенно возросло.

Разработка системы предполагает разработку и внедрение дополнительных функций, таких как отправка статистических данных и создание подсказок при составлении сценария. Анализ статистики поможет определить приоритетные направления по улучшению интерфейса.

Диапазон использования системы может быть расширен за счет первоначальной гибкости конструкции.

Литература

1. K. Aksyonov, D. Antipin, T. Afanaseva, I. Kalinin, I. Evdokimov, A. Shevchuk, A. Karavaev, O. Aksyonova, U. Chiryshe , “Testing of the speech recognition systems using Russian language models,” Yekaterinburg, Russian Federation, December 2018. [5th International Young Scientists Conference on Information Technologies, Telecommunications and Control Systems, ITTCS 2018].
2. K. Aksyonov, E. Bykov, O. Aksyonova, N. Goncharova, A. Nevolina, “Extension of the multi-agent resource conversion processes model: Implementation of agent coalitions,” pp. 593-597, 2016 [5th International Conference on Advances in Computing, Communications and Informatics].
3. K. Aksyonov, E. Bykov, E. Sysoletin, O. Aksyonova, A. Nevolina, “Integration of the Real-time Simulation Systems with the Automated Control System of an Enterprise,” pp. 871-875, 2015 [International Conference on Social Science, Management and Economics].
4. K. Aksyonov, I. Kalinin, E. Tabatchikova, U. Chiryshev, O. Aksyonova, E. Talancev, A. Tarasiev, V. Kanev. “Development of decision making software agent for efficiency indicators system of IT-specialists,” Yekaterinburg, Russian Federation, December 2018. [5th International Young Scientists Conference on Information Technologies, Telecommunications and Control Systems, ITTCS 2018].
5. “Исследование надежности распознавания речи системой GoogleVoiceSearch,” Cyberleninka.ru, 2018. [Электронныйресурс]. – URL: https://cyberleninka.ru/article/v/issledovanie-nadezh. (Дата обращения: 14.09.2019).
6. “Features of TWIN,” Twin24.ai, 2018. [Электронныйресурс]. – URL: https://twin24.ai/#features (Дата обращения: 14.09.2019.
7. A. Tarasiev, E. Talancev, K. Aksyonov, I. Kalinin, U. Chiryshev, O. Aksyonova, “Development of an Intelligent Automated System for Dialogue and Decision-Making in Real Time,” Bern, Switzerland, December 2018 [2nd European Conference on Electrical Engineering & Computer Science (EECS 2018)].
8. “Speech Kit Cloud”, Speech Kit Cloud, 2019. [Электронныйресурс]. – URL: https://tech.yandex.ru/speechkit/cloud/. (Дата обращения: 15.08.2018).
9. “SpeechKi”", Tech.yandex.ru, 2018. [Электронныйресурс]. – URL: https://tech.yandex.ru/speechkit/. (Дата обращения: 15.08.2018)..