Распознавание русской речи: состояние и перспективы

Хитров М.В.,
кандидат технических наук

За последние 2–3 десятилетия было множество публикаций, посвящённых созданию сначала в СССР, а затем и в России систем автоматического распознавания устной русской речи. Однако до сих пор на рынке отсутствует сколько-нибудь коммерчески состоятельный продукт. Почему же после стольких слов обещаний, демоверсий программ и т.д. так и не была реализована давнишняя мечта о создании автомата, реализующего столь естественную функцию, которая отличает человека разумного (HOMO SAPIENS) от остальных живых обитателей планеты Земля, по крайней мере для русской речи?

Обратимся сначала к зарубежной истории создания подобных средств для иностранных языков. На мой взгляд, первыми коммерчески успешными продуктами были программы для PC под Windows: Naturally Speaking компании Dragon и ViaVoice компании IBM, которые появились на рынке в конце 90-х годов прошлого столетия. *Большую роль в развитие технологий распознавания вносили государственные институты. Так, создание базы данных английского языка (американский вариант) было профинансировано из фондов Министерства обороны США в середине 80-х годов по заказу Управления перспективных оборонных проектов (Defense Advanced Research Projects Agency — DARPA). Тогда в США были созданы или начинали создаваться базы, которые у нас создаются только сейчас и без помощи государства (числительных, изолированных слов и слитной речи, в условиях без помех и телефонные). Вся территория США разбита на 21 диалектный район, каждый из которых был представлен 5 дикторами-мужчинами и 5 женщинами. В те же годы во Франции создавалась национальная речевая база данных.*

Решению проблемы распознавания спонтанной речи был посвящен японский национальный проект: «Spontaneous Speech Corpus and Processing Technology», который действовал в течение пяти лет 1999–2004, с бюджетом приблизительно в $10 млн US. В ходе этого проекта был собран и обработан крупномасштабный корпус спонтанной речи, Corpus of Spontaneous Japanese (CSJ), состоящий приблизительно из 7 млн слов с общим количеством речи 650 часов.

Создание таких баз данных — это трудоёмкая работа, рассчитанная на несколько лет и требующая постоянного обновления и дополнения. Отметим, что речевая база является одним из основных инструментов при создании современных систем распознавания речи.

На основе этих разработок созданы современные программы распознавания речи, которые вполне удовлетворительно работают, поддерживая наиболее распространённые в мире языки, к которым относятся: английский с различными версиями и диалектами (например, английский для США, Канады, Австралии и т.п.), испанский, немецкий, французский, китайский и др. В настоящее время права на программы компании Dragon принадлежат американской компании Nuance, которая является сейчас крупнейшим в мире игроком на рынке речевых технологий. Компании IBM, Microsoft продвигают собственные

продукты, которые соответствуют современному уровню, но, на мой взгляд, несколько уступают программам диктовки компании Nuance. Программы автоматического распознавания называют преобразователями речь — текст или иногда диктовочными блокнотами. Упомянутые программы работают вполне удовлетворительно: на нормативно правильном языке (имеется в виду носитель языка, без заметных дефектов речи) обеспечивается начальная точность распознавания от 85 до 90% с последующим повышением качества по мере адаптации программы к голосу пользователя (алгоритм адаптации реализован таким образом, чтобы подстройка программы к голосу происходила автоматически, без участия пользователя). К недостаткам существующих систем можно отнести их недостаточную робастность, т.е. недостаточное качество работы в неофисных условиях эксплуатации. До сих пор отмеченные системы нестабильны в условиях высоких окружающих шумов. Алгоритмы распознавания совершенствуются, но пока такие программы работают эффективно только в условиях офиса, с использованием специальных микрофонных гарнитур, обеспечивающих фиксированное расстояние между ртом говорящего и микрофоном, а также в ряде случаев обладающих средствами пассивного шумоподавления или шумоочистки.

Так почему же до сих пор нет подобных программ для русского языка?

Существует несколько причин как научных, так и коммерческих.

Начнем по порядку.

В семидесятые и особенно в восьмидесятые годы в СССР, где основным государственным языком был русский язык, активно велись разработки по речевой тематике многими научными коллективами, которые были сосредоточены в крупных отраслевых НИИ, таких как: НИИ Дальней связи (г. Ленинград), НИИ Автоматики (г. Москва), институт кибернетики (г. Киев) и др., в крупных вузах: МГУ, ЛГУ, НГУ и т.п. Большинство из речевых проектов было инициировано представителями правоохранительных органов, главным образом КГБ, а также Министерством обороны через уполномоченные органы АН СССР и союзных республик. Множество учёных было занято в этих проектах, которые не реже одного раза в 2 года собирались на научные симпозиумы, которые назывались АРСО (школа автоматического распознавания слуховых образов). Во время своего расцвета, например в 1984 г., в Новосибирском Академгородке собралось около 800 участников, что сравнимо по размерам с современными международными конференциями EUROSPEECH или INTERSPEECH, последняя из которых проходила в Антверпене и собрала около 1200 участников из разных частей всего мира (США, Великобритания, Япония, Китай, Франция, Германия и др.). К нашему стыду от России было всего несколько человек: не более 5, тогда как представителей, например, США было несколько сотен.

Результаты исследований советских времен соответствовали мировому уровню, но носили научно-прикладной характер, не ставящий перед собой цели успешного коммерческого использования этих результатов. Шло соревнование нескольких научных школ: ленинградская, новосибирская, грузинская, белорусская, украинская и др.: как на научном поприще, так и, особенно в последние годы этого ренессанса, в области создания макетных образцов систем

покомандного распознавания речи. Надо сказать, это соперничество дало свои плоды: были разработаны такие устройства, как «Марс» в Минске (школа Лобанова), «Барс» в Ленинграде (школа Галунова) и др. Эти устройства были вполне конкурентоспособными, но только не в СССР, где о рыночной экономике не могло быть и речи. Экспериментальная научная база советских учёных тех времён значительно отставала от уровня зарубежных коллег (это касалось и компьютерной базы), также отставала и элементная база, которую производила в те времена электронная промышленность СССР. Также в те времена только только начиналось освоение тех научных методов, на основе которых и создаются современные системы автоматического распознавания речи. Я имею в виду статистические методы построения акустических моделей, основанные на аппарате скрытых марковских моделей. Тот уровень компьютерной техники не позволял практическое использование подобных способов, по крайней мере в СССР, ввиду огромных вычислительных затрат, требуемых для полноценного обучение статистических моделей на больших корпусах речи того или иного языка.

Кроме отмеченных исторических причин, замедливших развитие науки в России, на отставание в области распознавания речи повлияли объективные трудности, связанные со спецификой славянских, а особенно русского языка. В области акустики наибольшую проблему для распознавания русской речи представляет необычайно сильная количественная и качественная редукция гласных безударных слогов, частично обусловленная свободным характером словесного ударения. Вкупе с низкой артикуляторной напряженностью это приводит к нейтрализации и «размазыванию» акустических свойств сегментов, особенно в спонтанной разговорной речи. С точки зрения грамматики и синтаксиса русский язык относится к синтетическим языкам со свободным порядком слов и богатой словоизменительной парадигмой, что существенно затрудняет языковое моделирование на основе «классической» n-граммной модели, поскольку требует использования чрезвычайно больших речевых корпусов для получения приемлемого числа реализаций всех входящих в словарь словоформ.

После исчезновения СССР и выхода из его состава союзных республик научное сообщество, вовлечённое в эту проблематику, было разъединено, а та «эйфория свободы», которой были преисполнены учёные и специалисты уже бывших союзных республик, многократно ускорила этот процесс всеобщего размежевания. Кроме того, в процессе либерализации, проводимой руководством России, фактически мгновенно поднялся «железный занавес» через который многие ведущие учёные и специалисты уехали вместе со своими наработками в ведущие исследовательские центры: США, Бельгии, Израиля, Австралии и т.д., способствуя тем самым значительному усилению позиций зарубежных конкурентов. В результате, в том числе и этих факторов, названные ранее компании (и не только), продвинулись в разработке своих продуктов и технологий в области распознавания речи.

Многие из перечисленных научных школ так и не оправились от утечки ведущих специалистов и всех прелестей либерализации российской экономики начала 90-х, но часть из них выжила вопреки тем безумным условиям существования научно-технических коллективов, но благодаря рыночным механизмам, которые были запущены в те времена. Имеется в виду, что в 90-е годы стало возможным основать и развивать эффективно работающие частные компании, в том числе и в сфере высоких технологий.

В конце 90-х годов недобрую службу в доверии частных лиц и, к сожалению, многих уважаемых государственных учреждений, таких как, например, МЧС, к речевым технологиям сыграл продукт под названием «Горыныч», который был выпущен на рынок (а не разработан, как многие считали; под разработкой я лично считаю серьезные мыслитель-

ные действия, приводящие к появлению продукта или технологии, обладающими качествами, позволяющими их использовать с высокой пользой) компанией, если не ошибаюсь, White. Программисты названной компании, взяв в качестве основы известную в то время на рынке программу распознавания Naturally Speaking компании Dragon, приспособленную для работы на английском языке, просто-напросто локализовали её на русский язык, ничего не меняя в так называемом «движке» (ядро системы распознавания, которое настроено на определённый язык). Пользователям предлагалось использовать программу «Горыныч» (остроумный синоним Dragon, собственно подчёркивающий её ориентированность на русский язык) для банального распознавания русскоязычных команд.

При этом наивному кандидату в покупатели показывались в качестве примеров речевые команды в исполнении специально подготовленных дикторов, которые удовлетворительно распознавались и на английском движке, что было аналогично действиям иллюзионистов или «напёрсточников» у станций метро в крупных городах. Насколько мне известно, достаточно большое количество копий так называемой программы распознавания речи «Горыныч» были приобретены государственной службой МЧС, специалисты которой верили в возможность её использования в реальных условиях работы сотрудников службы в местах происшествий. Бессовестные распространители этой программы обещали безупречную её работу даже в эпицентрах пожаров и прочих мест происшествий. Работа этих псевдоучёных, которые форманту от фонемы не способны отличить, на несколько лет отбила охоту обращать внимание на такие технологии со стороны серьёзных государственных заказчиков.

Сегодня имеется несколько российских и зарубежных компаний, которые приблизились к созданию коммерчески выгодных и практически полезных систем распознавания устной русской речи.

Приведу данные о состоянии системы распознавания речи в одной из них: Центр речевых технологий (ЦРТ), основной офис разработок которой находится в Санкт-Петербурге.

Специалисты ЦРТ разработали алгоритмы обучения акустических моделей фонем русского языка. Процесс обучения был выполнен на корпусах русской речи, собранных и имеющихся в распоряжении ЦРТ. Они представляют собой спектральные и динамические характеристики звуков речи, зависящие от соседних звуков. Многообразие сочетаний звуков в речи приводит к тому, что количество таких моделей может достигать нескольких тысяч. Набор моделей, использующихся в ЦРТ, может достигать 6 тысяч, причём количество их будет расти с ростом обучающей базы данных. Теперь, для того чтобы создать эталон распознаваемой команды, достаточно ввести её в текстовом виде с клавиатуры.

На основе нескольких миллионов словоформ, полученных в результате обработки текстов в русском сегменте Интернета, была создана языковая модель русской речи. В настоящее время заканчивается один из основных этапов работы — разработка декодера для системы распознавания, работающий с большим словарём. Декодер, работающий с ограниченным словарём, создан в начале года и успешно применяется в задачах поиска «ключевых слов» в потоке речи.

Как было отмечено выше, существеннной проблемой при практическом использовании систем распознавания речи является их низкая помехоустойчивость. В современных системах борьба с помехами ведётся на каждом этапе обработки речевого сигнала: используются направленные микрофоны, методы адаптивного подавления помех, адаптация акустических моделей к шумам *(рис. 1).*

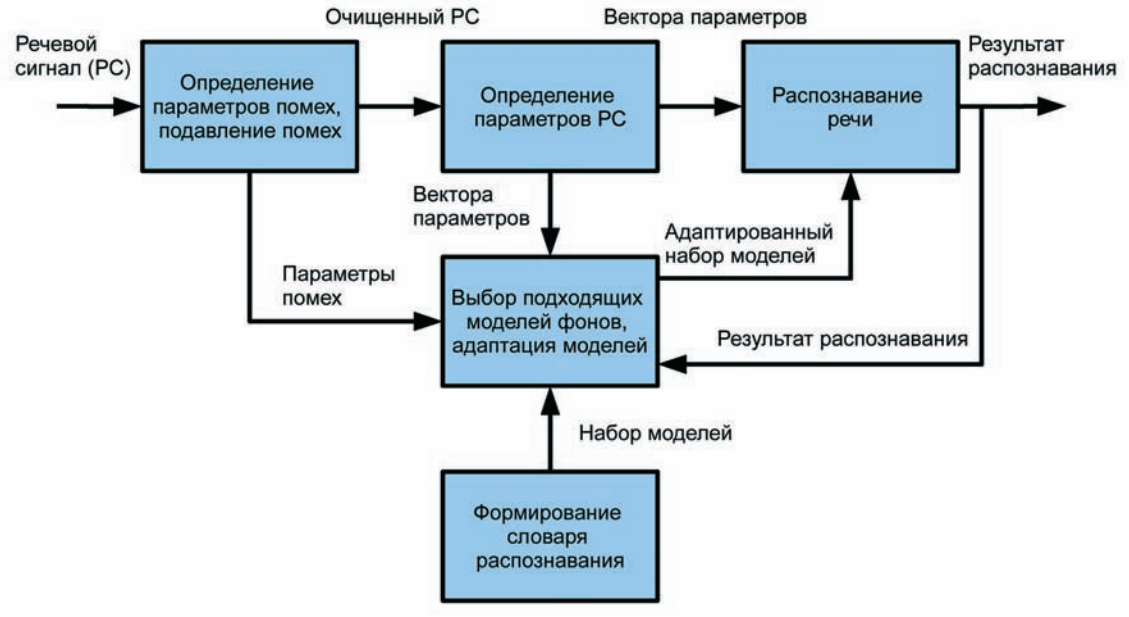

Рис. 1. Структрура современной системы распознавания

В настоящее время достигнуты значительные успехи по повышению помехоустойчивости распознавания для ограниченного класса стационарных помех (шум машины, шум улицы, шум самолёта и др.). Нерешённой проблемой остаётся распознавание в условиях нестационарных речеподобных помех. Даже при малом уровне таких помех качество распознавания речи существенно снижается.

Хитров Михаил,

кандидат технических наук.
Генеральный директор компании «Центр Речевых Технологий»