РЕЧЕВЫЕ ИНТЕРФЕЙСЫ ОТ SPEEREO

КОНСТАНТИН ЛАМИН
ОЛЕГ МАЛЕЕВ, К.Т.Н.
lamin@speereo.com

Российская компания Speereo выпустила первый в мире универсальный пульт дистанционного управления с использованием речевых команд. Разработчики рассказали о новинке, ответив нашему журналу на вопросы «Зачем это нужно?», «Что оно делает?», «Как это работает?», «Куда это будет развиваться?».

ЗАЧЕМ НУЖЕН РЕЧЕВОЙ ИНТЕРФЕЙС?

Речевой интерфейс (РИ) нужен в целом для упрощения жизни пользователям. Если конкретней, для повышения удобства, повышения степени интеллектуализации человеко-машинного диалога. Все это вполне счетные величины. Разработчики давно уже борются за такие параметры, как время на обучение пользователя, время отдачи команды, количество движений для отдачи команды, время на поиск нужного контрольного элемента. По всем этим параметрам введение речевого канала в подсистему интерфейса приводит к существенным улучшениям. Есть ряд применений, в которых речевое управление — настоящее спасение. Это ситуации, когда руки и зрение пользователя заняты важными «неинтерфейсными» задачами (вождение транспорта, визуальный осмотр, тонкие манипуляции, просмотр фильма). Если в этот момент требуется помощь компьютерной системы информации или робота-ассистента, без речевого диалога не обойтись. Эти простые соображения и дают нам основные сферы применения речевых интерфейсов:

- Бытовые сложные системы (бытовая техника, сервисные роботы и «умный дом»). Тут на первом месте скорость обучения пользователей, мобильность и упрощение интерфейсов.
- Системы поддержки деятельности людей, занятых вне офисных столов. Это водители, спасатели, ремонтники, военные, логисты, сборщики — всех не перечесть. Речевой канал здесь востребован в качестве возможности освободить руки и глаза. Важна и большая компактность решения.

Из понимания областей применения следует и набор требований к РИ:

- Безошибочность (количество ошибок на сотню слов, WER). Причем, для промышленных и бытовых применений WER нужно считать при различных окружающих шумах (соотношение сигнал/шум SNR).
- Количество различаемых команд в один момент времени. Чем сложнее объект управления и чем

◄ **КОНСТАНТИН ЛАМИН,** CEO Speereo

◄ **ОЛЕГ МАЛЕЕВ,** к. т. н., CTO Speereo

Компания «ЗАО «Титан — информационный сервис» / Speereo Software была основана в 1998 г. В 2001 г. ее специалистам удалось создать систему распознавания слитной английской речи, а в 2011 г. — слитной русской речи. С 2002 г. компания разрабатывает и продает продукты и решения, основанные на SSR (распознавание речи Speereo).
Компания — официальны поставщик Intel, имеет совместный грант Microsoft и Сколково, победитель конкурса инновационных проектов МО РФ, обладатель нескольких Best Software Award of the Year. С 2011 г. — резидент Сколково.

► Универсальный речевой пульт Speaky

меньше времени мы хотим учить пользователя, тем больше должен быть этот параметр.

- Антропоморфность. Это интегральный показатель, который отвечает за то, насколько интерфейс схож с человеческим общением. Очень широкая тема, но очевидно, что чем выше этот показатель, тем легче происходит обучение этому интерфейсу. Не следует путать с интуитивностью, которая лишь характеризует привычность и похожесть на уже известные пользователю интерфейсные системы.

ТАБЛИЦА 1. ЗАВИСИМОСТЬ ТОЧНОСТИ РАБОТЫ ASR SPEEREO ОТ УРОВНЯ ШУМА

SNR (db)	0	5	10	15	20	>50
WER	1,8	1,6	1,7	1,4	1,3	0,8

Примечание: языки — русский и английский, короткие слова (цифры) и длинные фразы — 600, 50 дикторов.

ТАБЛИЦА 2. ТОЧНОСТЬ РАБОТЫ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ В РАЗЛИЧНЫХ ШУМОВЫХ УСЛОВИЯХ

Система	SNR > 50 (94 фразы)	SNR 5-15 (82 фразы)	Среднее значение (176 фраз)
Apple Dictation	14,24	43,76	26,73
Bing Speech	11,73	36,12	22,05
Google API	6,64	30,47	16,72
Wit.ai	7,94	35,06	19,41
Baidu Deep Speech	6,56	19,06	11,85

Источник: https://gigaom.com/2014/12/18/baidu-claims-deep-learning-breakthrough-with-deep-speech/

► Министр обороны РФ Сергей Шойгу посетил стенд Сколково на неделе инноваций Министерства обороны в Алабино в августе 2014 г. Фото пресс-службы Сколково

ЧТО ДЕЛАЕТ РЕШЕНИЕ SPEEREO?

Мы разработали РИ, который позволяет распознавать речевые команды и синтезировать речевые сообщения. Система состоит из программной части (Automatic Speech Recognition, ASR, и Text To Speech, TTS), и аппаратной части — различного типа Acoustic Front End, AFE. Для разработчиков — это готовые блоки, которые можно встраивать в системы на этапе проектирования или на этапе апгрейда систем. ASR и TTS существуют как в виде «облачного» решения, так и в виде кода для «тонких» клиентов. Минимальные требования — 200 MIPS и 5 Мбайт. AFE существует на сегодня в виде серийного изделия — универсального речевого пульта Speaky, а также в виде тестовых прототипов и серийных изделий партнеров — автомобильного, OutDoor-гарнитуры, бытовой и игровой гарнитур. AFE может быть и чужим, лишь бы обеспечивал приемлемое качество сигнала. Реальное расстояние от микрофона до диктора, на котором обеспечивается нормальная работа без экстраординарных затрат на оборудование, составляет 10–50 см. Необходимо средствами AFE маркировать начало (обязательно) и конец (желательно) командной фразы. Для этого используются аппаратные кнопки, камеры, ларингофоны и пр. В мобильных версиях это позволяет экономить заряд батарей.

КАК ЭТО РАБОТАЕТ?

При минимальных требованиях по нагрузке вычислительной систе-

мы РИ Speereo работает с задержкой от конца фразы до выдачи результата не более 1,5 с. При этом мы добились очень высокого уровня по главным требованиям (табл. 1).

Для сравнения приведем редчайшую таблицу (табл. 2).

Как видим, даже сравнивать систему РИ Speereo и системы диктовки, получившие широкое распространение, не стоит. Разница на один–два порядка.

Количество одномоментно различаемых команд в нашей системе составляет от нескольких сот до 10 000. Управление мгновенным словарем отдано «на откуп» разработчику. Общий словарь системы не ограничен. Массив мгновенного словаря подается на вход системы динамически в виде текста. Это позволяет строить контекстнозависимые диалоговые системы.

Системы диктовки имеют ограниченный несколькими сотнями тысяч (до 2 млн.) слов мгновенный словарь. Он же — общий словарь. Добавлять новые слова может только разработчик системы.

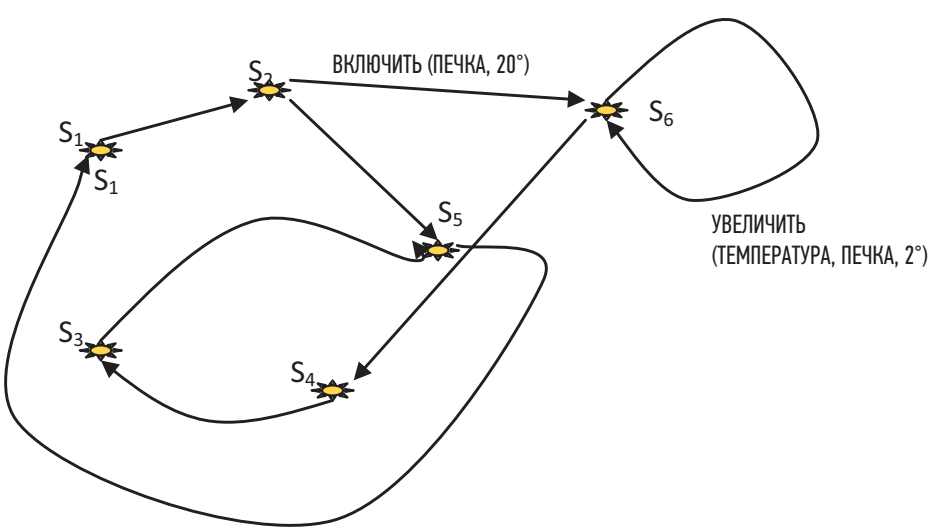
Ограничение в 10 000 фраз мгновенного словаря в нашей системе, тем не менее, позволяет строить интерфейсы для любых мыслимых объектов управления. Более того, в 10 000 фраз вполне укладываются все разумные варианты произнесения командных фраз в конкретный момент диалога. Для облегчения построения графа состояний и команд-переходов можно использовать распространенные грамматики. Таким образом, мы получаем систему, для работы с которой вообще не нужно обучать пользователя. Этот подход отличается от дилетантского заблуждения: «возьму систему диктовки, а потом разберу текст». Уровень ошибок систем диктовки и сложность систем «понимания» текста всегда ставит на таких планах крест. Немного улучшает ситуацию лишь очень трудозатратная система учета статистики поведения пользователей и семантической обработки, которая помогла, в частности, построить Google Voice Search и Apple SIRI, но провалилась в Google Glass, Apple TV и прочих проектах. Такие надстройки не поставляются внешним разработчикам, а их создание тянет на десятки миллионов долларов. Поэтому так мало внедрений систем диктовки вне обозначившихся узких ниш. Часть разуверившихся в диктовку разработчиков приходят к нам, как к альтернативному поставщику, и рассказывают очень похожие истории провала проектов.

◀ Прототип Speereo на неделе инноваций министерства обороны

Если еще учесть, что наша система дикторонезависима, устойчива к акценту, манере и темпу речи, не требует делать паузы между словами, то можно утверждать, что по параметру антропоморфности она находится на самом высоком современном уровне.

НЕМНОГО О ПЛАНАХ

Мы продолжим внедрение нашей системы в автомобильную, бытовую, специальную и промышленную электронику. Продолжим исследования по вычленению речевых сигналов из зашумленного потока, определения начала и конца команд. Мы также разворачиваем работы по повышению дальности от диктора до микрофона до нескольких метров, что позволит строить интерфейсные зоны в помещениях вообще без носимых устройств. Мы будем накапливать и делать доступными нашим потребителям тематические семантические сети, что облегчит построение «свободных» интерфейсов к целым классам техники. Продолжим работу и по увеличению размера мгновенного словаря. Все это вместе приближает нас к созданию близкого к идеалу речевого интерфейса. ⬣



▼ Ситуативная модель диалога схема