Речевые технологии для встраиваемых систем и мобильных устройств

О.Г. Малеев (ЗАО "Титан – Информационный сервис")

Рассмотрены характеристики технологий речевых интерфейсов (распознавание речевых команд и синтез речи), известных на рынке под торговой маркой Speereo. Показано, что речевые технологии с успехом могут применяться во встраиваемых и мобильных системах, построенных на широком классе платформ.

Для построения речевого интерфейса применяются вместе или по отдельности следующие технологии:

- распознавание речевых команд (ASR), обеспечивающие управление устройством с помощью голоса. ASR упрощенно можно представить как преобразователь речевого сигнала в текст или команду для управления устройством;

- синтез речи (TTS), позволяющий организовать речевой ответ устройства, его можно представить как преобразователь текста в речевой сигнал;

- компрессия речи, позволяющая эффективно хранить в небольшом объеме памяти речевые сигналы, а также передавать речь по каналам связи.

В том или ином виде данные технологии от различных поставщиков присутствуют на рынке достаточно давно, однако не все они обладают достаточными для коммерческого использования свойствами (особенно это касается систем распознавания речи). Если рассматривать сегмент встраиваемых и мобильных устройств, то здесь предложений от поставщиков еще меньше, несмотря на всеми понимаемую актуальность таких решений. Причина в том, что анализ и синтез речи – достаточно сложная научная и алгоритмическая задача, требующая больших ресурсов. Типичные приложения, ориентированные на Desktop PC, требуют для распознавания речи: Pentium 2,0 Гц, RAM 64 Мб с пропускной способностью шины памяти до 1,2 Гб/с; для синтеза речи: Pentium 2,0 ГГц, RAM 100...500 Мб. Для сравнения распространенный процессор Xscale 300 МГц обладает в 12 раз меньшей производительностью, чем Pentium 2,0 Гц, и намного более низкой пропускной способностью шины памяти – 64 Мб/с. Понятно, что стандартные решения не подходят для встраиваемых и мобильных устройств, нужны специальные подходы для снижения требований к CPU и памяти. Успех на этом сегменте рынка обеспечит технология, работающая на процессорах производительностью менее 100 MIPS и требующая не более 1 Мб памяти.

Другим важным моментом является легкость интеграции речевого интерфейса в конкретное устройство. Разработчику устройства, который, скорее всего, не является специалистом в области речевых технологий, должны быть предоставлены библиотеки и документация (SDK), с помощью которых он мог бы легко интегрировать речевой интерфейс в разрабатываемое устройство. Поставляемое решение должно поддерживать устройства с различными операционными системами и даже без них. По экономическим причинам желательно, чтобы применение речевых технологий не требовало установки дополнительных чипов, так как это ведет за собой внесение изменений в дизайн печатной платы устройства.

Конечно, есть еще специфические возможности технологии, определяющие качество работы речевого интерфейса. Рассмотрим характеристики технологии распознавания речи на примере технологии, разработанной компанией "Титан-Информационный сервис", известной на рынке под маркой Speereo.

Точность работы системы. Никому не понравится, если устройство будет понимать его неправильно. Если технология не обладает достаточной точностью (≥ 95%), то это лишь игрушка или прототип, но не коммерческая система. Speereo ASR обеспечивает беспрецедентную точность распознавания. В проводимых экспериментах система в 99,9% случаев понимала произносимые человеком команды.

Второй важной характеристикой технологии является возможность работы системы с любым диктором. Система не должна требовать подстройки к конкретному человеку. В ряде случаев такую подстройку выполнить просто невозможно (например, в публичных информационных киосках, когда диктор заранее неизвестен). Однако и в других ситуациях подстройка обременительна для пользователя. Кроме того, дикторонезависимость означает определенный показатель уровня технологии и гарантию того, что система будет понимать пользователя, даже если он простужен и у него изменился голос.

Speereo ASR изначально разрабатывалась как дикторонезависимая и в английском варианте уже опробована на всех континентах. Система понимает даже шотландцев, индийцев, корейцев и других пользователей, для которых английский язык не родной. По этому критерию технология уникальна.

Существует еще много показателей, по которым оцениваются подобные системы. Например, гибкость изменения речевых команд. Speereo ASR позволяет менять речевые команды динамически во время исполнения программы. Словарь системы содержит более 100 тыс. слов. Более того, благодаря специальному модулю, можно распознавать слова, не входившие изначально в словарь системы. Другая важная характеристика – робастность по шуму. Speereo ASR позволяет уверенно понимать речевые команды в условиях повышенных шумов, в том числе автомобильных.

Ниже приводятся результаты тестов Speereo ASR в различных условиях.

Тест 1. Распознавание длинных фраз. Условия теста: 1680 произнесений, словарь содержит 626 уникальных фраз, язык – английский. Точность распознавания – 99,9%.

Тест 2. Распознавание коротких слов. Условия теста: цифровой словарь, 1543 произнесения, словарь содержит 11 слов. Язык – английский, точность распознавания – 99,2%. Язык – русский, точность распознавания – 98,5%.

Тест 3. Робастность по шуму. Параметры теста: 1543 произнесения (база теста 2 была протестирована в условиях шума). Результаты представлены в таблице.

Тест 4. Точность системы в условиях автомобильных шумов. Условия теста: 1680 произнесений (база аналогична тесту 2). Распознавание проводится в автомобиле, который двигается по трассе со скоростью 120 км/ч с открытыми передними окнами. Язык – английский, точность распознавания – 97,6%.

Таблица. Точность Speereo ASR в условиях шума

Сигнал/Шум, dB	0	5	10	15	20	Clear
Точность, %	98,2	98,4	98,3	98,6	98,7	99,2

Speereo ASR, как видно из приведенных тестов, демонстрирует исключительно высокие результаты по точности, в том числе в условиях повышенных шумов.

Speereo ASR в настоящее время может работать на большинстве широко распространенных процессорах для мобильных и встроенных устройств таких, как SHx, TMPR39XX, NEC VR4122, MIPS, ARM, Xscale. Speereo ASR работает с CPU производительностью от 40 MIPS (рекомендуется 80 MIPS) и памятью от 700 Кб, что является уникальными показателями для такого типа технологий.

Многие разработчики устройств думают, исходя из сложности задачи распознавания, что интеграция речевого интерфейса с их устройством потребует от них больших усилий. Однако это не так. Разработчик при лицензировании библиотек для распознавания получает доступ к использованию простого API, с помощью которого можно проектировать речевые интерфейсы без специальных знаний в области речевых технологий. API несколько отличается для разных ОС в плане деталей реализации в связи с необходимостью учета особенностей каждой платформы (сейчас существуют версии для Windows CE, Symbian и для устройств без ОС). Однако основные принципы остаются неизменными.

Рассмотрим пример API Speereo ASR для Windows CE. Speereo ASR берет на себя все необходимое управление микрофоном и динамиком устройства, освобождая разработчика от программирования устройств ввода/вывода речевого сигнала.

В работе Speereo ASR могут быть выделены три основных этапа: регистрация приложения в ASR (для многозадачных ОС); отправка списка речевых команд ASR; при произнесении фразы (команды) ASR определяет наиболее вероятную фразу из списка всех фраз и передает ее ID приложению.

Разработчику не нужно отслеживать момент произнесения фразы. Все, что необходимо – это обработать сообщение Speereo ASR, содержащее ID команды, произнесенное пользователем.

Таким образом, благодаря Speereo ASR встроить речевой интерфейс в устройство становится очень просто.

Логично, чтобы устройство, понимающее речевые команды, выдавало ответные сообщения пользователю тоже с помощью голоса. Достигается это с помощью специальной технологии синтеза речи (TTS). В случае интегрирования TTS в свое устройство разработчик должен определить, нужен ему синтез предопределенных фраз (например, стандартные ответы на действия пользователя) либо свободного текста (например, чтение новостей из Internet). В случае предопределенных фраз возможно использование целословного синтезатора. Если же фраз много либо они меняются, необходимо применение фонемного синтезатора. Целословный синтезатор содержит в своей базе записанные слова, из которых может с помощью некоторых алгоритмов составлять необходимые фразы.

Преимуществом таких синтезаторов является легкая смена голоса и языка синтеза, а также небольшие требования по памяти (в случае небольшой базы слов – до 3000 ед.) и ресурсам процессора. Недостатком, кроме ограничений размера словаря, можно назвать невозможность придания интонаций генерируемой фразе. Фонемные синтезаторы содержат в своей базе более мелкие частицы речи и могут генерировать любые фразы из большого словаря (>100 тыс. слов) с любой интонацией. Это более универсальное решение, требующее больших ресурсов CPU и памяти. Также здесь более сложно добавить новый голос и язык.

Для примера приведем характеристики систем Speereo TTS.

Целословный TTS. Известный на этапе разработки системы словарь (до 2...3 тыс. слов). CPU от 40 MIPS, RAM от 0,5 Мб, требует произнесения диктором всех слов словаря. Язык любой. Время кастомизации для конкретного устройства 1...2 недели в зависимости от словаря.

Фонемный TTS. Могут использоваться большие словари (более 100 тыс. слов). CPU от 80 MIPS, RAM от 2 Мб, не требует настройки на словарь. Сейчас есть поддержка английского и испанского языка. Время добавления в синтезатор нового языка 3 месяца.

В целословном синтезаторе для снижения занимаемой памяти применяется компрессия речи. Алгоритмы компрессии речи имеют и самостоятельное значение для записи каких-либо речевых комментариев пользователя либо предопределенных фраз в памяти устройства.

Компанией "Титан-информационный сервис" разработан алгоритм, позволяющий записать 1 минуту речевого сигнала примерно в 10,25 Кб, т. е. в 1 Мб памяти можно записать более 1,5 часов речевого сигнала. Алгоритм компрессии/декомпрессии Speereo в PB требует процессора производительностью 60 MIPS и 200 Кб памяти. Только алгоритм декомпрессии Speereo в PB требует процессора производительностью 40 MIPS и 200 KB памяти.

Таким образом, речевые технологии с успехом могут применяться во встраиваемых приложениях и мобильных системах, построенных на широком классе платформ. Внедрение речевого интерфейса с использованием Speech SDK фирмы "Титан-информационный сервис" не требует больших усилий со стороны разработчиков устройства. Не нужно использовать дополнительные чипы и владеть специальными знаниями в области технологий. Рассмотренные решения требуют небольших ресурсов, а использование речевого интерфейса позволяет обеспечить новый уровень в удобстве управления встроенными и мобильными устройствами.

Малеев Олег Геннадиевич *– канд. техн. наук, директор по научно-исследовательским разработкам ЗАО "Титан – Информационный сервис". Контактный телефон (812) 327-43-18. maleev@speereo.com Http://www.speereo.com*