ПЕРСПЕКТИВНАЯ РЕЧЕВАЯ ТЕХНОЛОГИЯ «ИНТЕРФЕЙС БЕЗМОЛВНОГО ДОСТУПА»

А.А. Кравцов

А.А. Кравцов, начальник научно-исследовательской части Московского государственного лингвистического университета, канд. техн. наук, профессор, Почетный работник науки и техники РФ
kravtsov@linguanet.ru

В статье приведены результаты анализа «революционной» речевой технологии на основе Интерфейсов Безмолвного Доступа (Silent Speech Interfaces — SSI). Рассматриваются научные принципы, положенные в основу новой технологии, потенциал применения новой технологии для преодоления основных недостатков, присущих традиционным речевым интерфейсам, концепции и механизмы использования SSI для различных вариантов коммуникации.

The article deals with the results of the analysis of revolutionary speech technology based on the Silent Speech Interface (SSI). The scientific principles taken as a principle new technology, potential of application of new technology for overcoming of the basic lacks inherent in traditional speech interfaces, concepts and mechanisms of use SSI for various variants of communications are considered.

В последнее десятилетие благодаря стремительному прогрессу в области микропроцессорной техники были достигнуты значительные успехи в создании новых речевых технологий. Это привело к расширению спектра коммуникационных услуг, таких как информационно-поисковые системы, call-центры, голосовое управление мобильными телефонами и автомобильными навигационными системами, транскрайберы и персональные переводчики, а также к применению речевых технологий в сфере безопасности. Поистине «революционным прорывом» в данной области, по мнению ведущих мировых специалистов, можно считать разработку технологии так называемых Интерфейсов Безмолвного Доступа (Silent Speech Interfaces — SSI).

Потребность в разработке данной технологии обусловлена тем, что существующие речевые интерфейсы, базирующиеся на традиционных акустических речевых сигналах, имеют ряд существенных ограничений. Во-первых, акустические сигналы, передаваемые через воздух, подвержены искажениям из-за шумов естественного и искусственного происхождения. Надежных систем обработки речи, которые бы безукоризненно функционировали на объектах массовых коммуникаций в условиях повышенного шумового фона (в переполненных ресторанах, аэропортах и других общественных местах), несмотря на титанические усилия разработчиков, по-прежнему не существует. Во-вторых, традиционные речевые интерфейсы требуют достаточно чёткой и внятной речи, что имеет целый ряд недостатков, в частности, при осуществлении речевой коммуникации в общественном месте возникает угроза конфиденциальности речевого сообщения.

Учеными Западной Европы (Германия, Франция), США и Японии в начале прошлого десятилетия был предложен новый подход к решению вышеуказанной проблемы — использование при коммуникации системы Интерфейсов Безмолвного Доступа. **Идея новой технологии заключается в том, чтобы частично или полностью исключить из процесса коммуникации слышимый акустический сигнал.** Эти системы базируются на получении и цифровой обработке речевых сигналов на ранней стадии артикулирования речевых сообщений коммуникантами. Благодаря этому пользователи системы приобретают возможность совершать коммуникацию практически «безмолвно», то есть без произнесения каких-либо звуков.

Концепция механизма действия системы SSI базируется на снятии параметров движений языка и губ в начальных процессах коммуникации с помощью комплекса ультразвуковых и оптических устройств, а также аппаратуры высокоскоростной обработки визуальных данных и фонемных компонент сгенерированных речевых сигналов (**рис. 1**).

Возможность осуществления коммуникации без использования речевого сигнала основывается на комплексе знаний о процессах речеобразования и речевосприятия. Известно, что речевые сигналы для передачи семантической (смысловой) информации путем устной речи первоначально генерируются на самых ранних этапах коммуникации — в момент задействования человеком своего артикуляционного аппарата, то есть еще до того, как речевое сообщение появится в воздухе. Изучение тождественных по типу звуков различных языков показало, что произносятся они с помощью достаточно тождественных, но в то же время несколько различных, присущих разным языкам приемов, которые образуют так называемую артикуляторную (или артикуляционную) базу данного языка. Артикуляторная база реализуется как на сегментном, так и на супрасегментном уровнях. Под сегментами понимаются звуки речи или их сочетания, располагающиеся последовательно в речевом потоке. К супрасегментным единицам относятся ударение во всех его видах, годика, темп, громкость. Они отличаются от сегментных единиц тем, что не могут существовать сами по себе, без звуков (сегментов), а как бы накладываются на них. Кроме того, установлено, что, несмотря на то, что существуют определенные мутации,

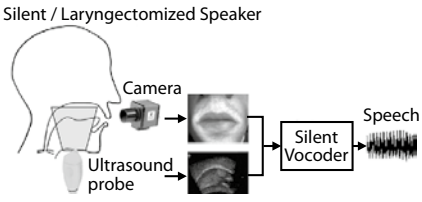

Рис. 1. Механизм действия системы SSI

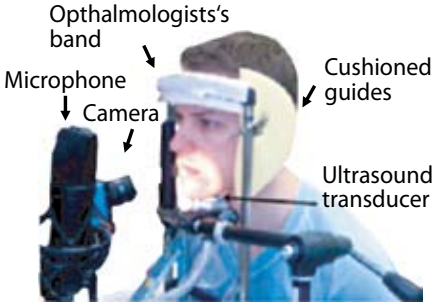

Рис. 2. Комплекс аппаратуры SSI, осуществляющий анализ речевых сигналов на ранней стадии артикуляции и оптических изображений изменения лицевых мышц

тем не менее, большинство людей обладают мимикой лица, которая функционально связана с артикуляторной базой языка, причем у большинства людей, говорящих на одном языке, она практически одинаковая. Таким образом, существует возможность до появления слышимых звуков эти артикуляционные сигналы зафиксировать с помощью сенсоров различного физического происхождения и после преобразований передать от источника к получателю для дальнейшего считывания, обработки и интерпретации.

История зарождения технологии непродолжительна, в частности, в 2002 году было экспериментально доказано, что электромиографические сигналы (от артикуляционных лицевых мышц) содержат достаточное количество информации, чтобы обеспечить с требуемой точностью распознавание небольшого набора отдельных слов; в 2006 году с помощью электромиографических сигналов было осуществлено распознавание слов и идентификация диктора в процессе речи.

Разработчиками Интерфейсов Безмолвного Доступа было доказано, что электромиографические сигналы содержат достаточно информации, чтобы эффективно распознавать фонемные единицы на основе анализа электрической активности мышц лица, а разработанные программы способны на основе получаемых визуальных изображений практически в реальном масштабе времени сегментировать лица, фиксировать изменения положения точек (мимики) лица и интерпретировать их в готовые к произнесению звуки речи. При этом исследователями было установлено, что само по себе видеораспознавание (без знаний речевой семантики), способно обеспечить точность распознавания до 60% (в частности для определения звуков американского английского языка).

К 2010 году французскими и немецкими исследователями были уже практически разработаны алгоритмы безмолвного распознавания речи, основанные на комплексном анализе ультразвуковых сигналов и оптических изображений лица, что открывает путь для распознавания не отдельных слов, а обширных словарных баз. Были разработаны образцы, которые обеспечивают процедуру распознавания речи, базирующуюся на словарном запасе в 3000 слов, с точностью распознавания фонем — 83%, слов — 62%. Экспериментально было установлено, что на основе комплексного анализа на имеющейся в настоящий момент видео- и аудио-аппаратуре, может быть обеспечено повышение точности распознавания речи более чем на 30% по сравнению с распознаванием только на основе фонемного анализа (**рис. 2**).

Механизм действия системы SSI базируется на том, что речевое сообщение от коммуниканта-источника на этапе его генерирования артикулационным аппаратом фиксируется различными неакустичекими датчиками и после цифрового преобразования передается напрямую через телекоммуникационное цифровое устройство коммуниканту (коммуникатам) — получателю (получателям) сообщения. где оно декодируется и синтезируется с помощью соответствующей аппаратуры со специальным программным обеспечением в смысловое речевое сообщение, аналогичное сгенерированному источником.

В проводимых экспериментах фиксирование речевого сообщения на ранних стадиях осуществлялось семью типами датчиков: 1) фиксирующими изменения артикуляционного аппарата с помощью электромагнитных артикуляционных сенсоров (Electromagnetic Articulography sensors); 2) снимающими в реальном масштабе времени параметры вокального тракта путем ультразвукового и визуального зондирования языка и губ; 3) фиксирующими и осуществляющими цифровое преобразование неслышимых слов (без вибрации голосовых связок) от микрофона Non-Audible Murmur microphone (работает по принципу медицинского стетоскопа), закрепляемыми на липучках за ушами человека (**рис. 3**); 4) измеряющими и анализирующими гортанную деятельность через вводимые в мышцы игольчатые и/или кожные электроды (так называемые электромиографические маски); 5) фиксирующими электромиографические сигналы артикуляционных мускулов и гортани на лицевой поверхности; 6) снимающими и преобразовывающими сигналы от электроэнцефалографических датчиков (**рис. 3**); 7) фиксирующих и преобразующих сигналы от электродов, имплантированных в кору головного мозга.

Датчики, фиксирующие электромиографические сигналы, изготовляются в различных вариантах: в виде: пластин, закрепленных на липучках на шее (**рис. 4**); электродов, замаскированных под ювелирные изделия (**рис. 5**).

Система SSI обладает большим потенциалом для преодоления основных недостатков сегодняшних традиционных речевых интерфейсов, таких как: а) ограничение надежности распознавания речевого сигнала при наличии фонового шума, б) риск доступа посторонних лиц к частной и конфиденциальной речевой информации, в) беспокойство окружающих, г) ущемление возможностей в устной коммуникации лиц с ограниченными возможностями по слуху и речи. Видео-распознавание речи способно существенно скоррелировать аудио-помехи и тем самым значительно увеличить точность самого распознавания. Зарубежные и отечественные эксперты единодушны во мнении, что на сегодняшний день создание Интерфейсов Безмолвного Доступа является самым перспективным направлением в области развития речевых коммуникаций.

Рис. 3. Датчик, фиксирующий и осуществляющий цифровое преобразование неслышимых слов (без вибрации голосовых связок)

Рис. 4. Датчик, фиксирующий изменения артикуляционного аппарата

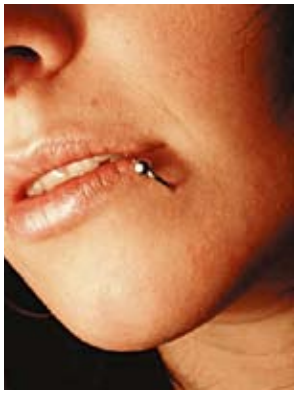

Рис. 5. Датчики фиксации изменения артикуляционного аппарата, закамуфлированные под ювелирные изделия