МНОГОМОДАЛЬНЫЕ ИНТЕРФЕЙСЫ: ОСНОВНЫЕ ПРИНЦИПЫ И КОГНИТИВНЫЕ АСПЕКТЫ♦

А. Л. Ронжин, А. А. Карпов

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178

<{ronzhin,karpov}@iias.spb.su>

УДК 681.3

Ронжин А. Л., Карпов А. А. **Многомодальные интерфейсы: основные принципы и когнитивные аспекты** // Труды СПИИРАН. Вып. 3, т. 1. — СПб.: Наука, 2006.

Аннотация. *При проектировании многомодальных систем за основу берутся способы естественной коммуникации между людьми, а также моделируется поведение человека в аналогичной ситуации. Многодомодальность позволяет вести более эффективный диалог за счет дублирования передаваемой информации по разным информационным каналам и выбирать (замещать) доступные для пользователя модальности. Выделяются два типа модальностей: входные и выходные. Входные отвечают за восприятие информационных потоков, идущих от человека (речь, звуки, движения тела, рукописный текст и др.). Выходные модальности обеспечивают пользователя необходимой информацией о событиях, происходящих внутри системы, и поступающих сигналах. Знания о механизмах восприятия и обработки информации человеком, моделировании когнитивных и поведенческих процессов позволяют разработать многомодальные интерфейсы, обеспечивающие привычные и естественные для пользователя способы взаимодействия.* — Библ. 15 назв.

UDC 681.3

Ronzhin A. L., Karpov A. A. **Multimodal Interfaces: Main Principles and Cognitive Aspects** // SPIIRAS Proceedings. Issue 3, vol. 1. — SPb.: Nauka, 2006.

Abstract. *The development of multimodal systems is based on natural forms of inter-human communication and modeling of human's behavior at the same situation. Multimodality provides more effective dialogue by duplication of the transmitted information via different channels and possibility of choosing the modalities accessible for a user. Two types of the modalities are discussed: input and output. The input modalities response for perception of information streams going from a human (speech, sounds, body motions, handwriting, etc.). The output modalities inform a user about the events taking place in the system and coming signals. Knowledge about mechanisms of human perception and information processing, modeling of cognitive and behavior processes allow to develop the multimodal interfaces, which provide usual and natural ways of the human-computer interaction.* — Bibl. 15 items.

1. Введение

Разработка средств эффективного взаимодействия человека с компьютером сегодня является одним из приоритетных направлений развития искусственного интеллекта и информатики в целом. Это связано с тем, что уже сейчас вычислительная техника не используются в полной мере из-за отсутствия полноценного, привычного человеку, интерфейса для взаимодействия пользователя с компьютером. Отсутствие решения этой проблемы сдерживает развитие многочисленных компьютерных прикладных систем в телекоммуникации, медицине, образовании и повседневной жизни, поскольку вся современная техника и различные сетевые сервисы используют автоматизированные средства управления и обработки информации.

♦Данное исследование проводится в рамках Европейской научной сети SIMILAR NoE «The European taskforce creating human-machine interfaces SIMILAR to human-human communication» FP6-IST № 507609 при финансовой поддержке ЕС.

Сегодня большинство компьютерных приложений используют графический пользовательский интерфейс, который обеспечивает весьма ограниченный способ взаимодействия: печать с помощью клавиатуры, управление виртуальными объектами курсором мыши и отображение визуальной информации в виде текста и изображений на экране монитора. Такой способ общения заставляет пользователей адаптироваться к компьютеру и учиться виртуальному способу общения. В результате пользователь вынужден ограничивать свои чувства и способы взаимодействия для того, чтобы получить доступ к компьютерному миру.

Для решения глобальной проблемы человеко-машинного взаимодействия необходимо использовать дополнительные виды каналов передачи информации (речь, артикуляция губ, жесты, направление взгляда и т.д.). Такой способ взаимодействия получил название «многомодальное взаимодействие», которое реализуется путем многомодальных интерфейсов. Такие интерфейсы свойственны межчеловеческому общению. Здесь мы сами выбираем, какой канал, для передачи какого типа информации нам наиболее удобно использовать в данный момент. Такие интерфейсы позволяют обеспечить наиболее эффективное и естественное для человека взаимодействие с различными автоматизированными средствами управления и коммуникации. В многомодальных системах информация от аудио, видео, тактильных и других коммуникативных каналов непрерывно обрабатывается, создавая реальное или виртуальное окружение, позволяющее удовлетворить желания пользователя, и оперативно адаптироваться к контексту.

В настоящее время за рубежом многомодальные интерфейсы уже используются в некоторых прикладных областях: картографических системах, системах виртуальной реальности, медицинских системах, робототехнике, web-приложениях, и т.д. Помимо этого, многомодальный интерфейс может быть полезен в мобильных устройствах, где использование обычной клавиатуры невозможно. В карманных персональных компьютерах сейчас используются системы распознавания рукописного текста. Комбинирование таких систем с голосовым вводом позволит обмениваться информацией с пользователем более эффективно [1]. Также использование многомодальных интерфейсов актуально в смартфонах (умных телефонах), в которых в настоящее время возможен раздельный ввод с помощью голоса, неэргономичной клавиатуры или сенсорного экрана. Совместное использование нескольких коммуникативных каналов позволит пользователю более оперативно и надежно обмениваться информацией с такими устройствами.

В России научные исследования по данному направлению начались совсем недавно, и их успешная реализация усложняется тем, что необходимо объединять усилия различных исследовательских групп, занимающихся отдельно обработкой речи, видеоизображений, почерка, и т.д. в различных научно-исследовательских институтах. С 2003 года группа речевой информатики СПИИРАН ведет фундаментальные и прикладные работы по многомодальным интерфейсам в рамках Европейского научного сообщества SIMILAR, финансируемого ЕС по программе FP6.

Во втором разделе представлена терминология, используемая в данной области, рассматриваются основные достоинства многомодальных интерфейсов и дана их классификацию по принципу используемых модальностей. Выделяются два типа модальностей: входящие и выходящие. Их особенности и способы применения обсуждаются в третьем разделе. Четвертый раздел посвя-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

щен вопросам когнитивной науки, процессам восприятия и передачи информации, присущих человеку во время коммуникации между людьми, Моделирование многомодального поведения человека является основой для проектирования интеллектуальных систем естественного взаимодействия. Наиболее частые заблуждения, которые возникают при разработке многомодальных интерфейсов, обсуждаются на протяжении всего раздела и сформулированы в конце. В заключение статьи очерчено текущее состояние развития многомодальных систем, наиболее актуальные научные вопросы и перспективные приложения.

2. Основные принципы многомодальных интерфейсов

В области многомодальных исследований используются специальные термины и положения, которые следует сразу пояснить. Многомодальное человеко-машинное взаимодействие опирается на ряд принципов [3]:

1) пользователь управляет компьютером, используя несколько физических устройств (клавиатура, мышка, микрофон, видеокамера и т.д.);
2) для коммуникации с компьютером пользователь активизирует движение ряда своих мышц (голосового тракта, рук, глаз и т.д.);
3) информация, передаваемая компьютерными устройствами ввода, может быть обработана на различных уровнях абстракции, обеспечивая различные уровни понимания намерения пользователя;
4) компьютер взаимодействует с пользователем, используя несколько устройств вывода (дисплей, динамики и т.д.);
5) по этим устройствам вывода компьютер может передавать заранее подготовленные данные (файлы с изображениями, аудио файлы и т.д.) или же динамически генерируемые данные (например, генерация текста, графики, синтез речи и т.д.).

Таким образом, компьютерная система может использовать несколько информационных каналов (чувств пользователя) для ввода и вывода. Существует пять человеческих чувств (слух, зрение, вкус, осязание, обоняние), и термин «модальность» используется в контексте этих сенсорных способов восприятия информации. Например, с перьевым вводом связано несколько модальностей, таких как: рисование, рукописный ввод и жесты для ввода информации в компьютер; а с экраном монитора связаны: текст, графика, изображения, видео.

Так как речь по своей природе многомодальна, то можно говорить также и о многомодальных аспектах распознавания речи. Люди сопровождают речь также и невербальными способами выражения информации, включая выражение лица, направление взгляда, движения губ. Многомодальные речевые системы (аудиовизуальные) являются попыткой достичь той же простоты коммуникации, соединяя автоматическое распознавание речи с другими невербальными средствами, а также интегрируя невербальные средства с синтезом речи для улучшения метода вывода информации в многомодальном приложении (например, виртуальная говорящая голова).

По сравнению с традиционными компьютерными интерфейсами на основе клавиатуры и мыши или одномодальными интерфейсами, многомодальные системы обеспечивают более гибкое использование входных потоков информации. Это дает возможность пользователю выбирать наиболее удобный способ передачи различной входной информации, так как некоторые комбинации модальностей для передачи информации хорошо подходят для отдельных ситуаций и прикладных задач, но хуже или даже совсем неприменимы для других.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Среди основных преимуществ, которые позволяет получить применение многомодальных интерфейсов, можно выделить следующие [9]:

- синергизм модальностей. Синергизм модальностей может достигаться как на входных, так и на выходных модальностях. На входе использование нескольких модальностей может привести к повышению точности интерпретации фразы как, например, комбинирование распознавания речи с чтением по губам в условиях окружающего акустического шума. На выходе же объединение модальностей повышает информативность и естественность оповещения пользователя;
- имея несколько модальностей можно получить менее сложную и более функциональную систему. Например, указание на графический объект проще выразить при помощи указки, чем речевой командой, а команду проще сказать, чем выбирать из меню;
- новые приложения. Некоторые задачи сложно или даже невозможно выполнить при использовании только одной модальности. Например, интерактивное телевидение проще использовать в речевом диалоге, чем при помощи кнопок пульта управления или же взаимодействуя с системой меню. Однако, в текстовых редакторах удобнее пользоваться клавиатурой;
- свобода выбора. Хотя одна и та же задача может быть выполнена настолько же эффективно при помощи различных комбинаций модальностей, пользователи могут иметь другие, индивидуальные предпочтения и выбирать более удобные для них модальности;
- естественность. Для пользователя является более естественным, если он использует для взаимодействия с компьютером те же самые средства и каналы, что и при общении с людьми;
- адаптация к окружению, при которой может происходить переключение между используемыми входными модальностями в зависимости от внешних условий (шум, свет и т.д.).

Многомодальные интерфейсы принципиально отличаются от существующих графических пользовательских интерфейсов по нескольким показателям. Во-первых, графические пользовательские интерфейсы обычно предполагают, что есть один поток событий, который обрабатываются последовательно. Например, большинство графических интерфейсов игнорируют ввод с клавиатуры, когда нажата кнопка мыши. В противоположность этому, многомодальные интерфейсы обрабатывают непрерывный ввод параллельных потоков информации.

Во-вторых, графические интерфейсы предполагают, что основные действия, такие как выделение некоторого объекта, являются атомарными и однозначно выраженными событиями. В отличие от этого, многомодальные интерфейсы обрабатывают входные данные при помощи различных технологий распознавания и основываются на вероятностных методах и говорить о каком-то действии или событии можно только с учетом его вероятности.

В-третьих, графические интерфейсы пользователя обычно разрабатываются отдельно от программного обеспечения приложений, которым они управляют, хотя компоненты интерфейса располагаются на одном компьютере. Интерфейсы же, основанные на технологиях распознавания, предъявляют большие вычислительные требования, а также требования к объему компьютерной памяти, что вынуждает распределять такой интерфейс по разным компьютерам в сети, каждый из которых содержит различные распознаватели и базы данных.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Например, мобильные телефоны или карманные компьютеры могут вычислять признаки речевого сигнала и передавать их распознавателю, который расположен на сервере. Наконец, интерфейсы, обрабатывающие несколько потоков данных, требуют разработки различных временных ограничений и пороговых значений для объединения модальностей, т.е. создания архитектур чувствительных ко времени.

Многомодальные интерфейсы обеспечивают выбор модальностей для передачи различных типов информации и совместное использование модальностей. Так как конкретные модальности хорошо подходят в некоторых ситуациях и совершенно не годятся в других, то выбор модальностей — это важная задача при разработке многомодальных систем. Новейшие системы становятся более сложными и многофункциональными, и одномодальные системы уже не позволяют пользователям эффективно взаимодействовать со всеми задачами и приложениями.

Многомодальные интерфейсы обеспечивают адаптивность, необходимую для приспособления к постоянно меняющимся условиям эксплуатации. В частности, интерфейсы, использующие речевой ввод, письменный, тактильный ввод хорошо подходят для мобильных систем, в которых пользователи предпочитают время от времени изменять использование модальностей в зависимости от смены окружающей обстановки (например, от уровня освещенности или акустического шума). Есть также случаи временной недееспособности, когда человек не способен использовать отдельные виды ввода в течение некоторого периода времени. Например, пользователь при управлении автомобилем не может использовать интерфейс с компьютером, основанный на жестовом вводе или направлении взгляда, а вот речевой интерфейс в данных условиях является наиболее предпочтительным. Таким образом, многомодальные интерфейсы позволяют осуществлять выбор используемых модальностей и переключение между ними при изменении окружающей обстановки. Они удовлетворяют почти всем предпочтениям пользователя при взаимодействии с компьютерными системами.

Первой многомодальной системой принято считать систему “Put That There”, созданную в США в 1980-х годах [5]. Со времени появления этой первой демонстрационной многомодальной системы, которая обрабатывала речь параллельно с указаниями на сенсорной панели, было создано множество многомодальных систем. Один из способов классификации многомодальных систем по типам задач представлен на рис. 1. На верхнем уровне классификации многомодальные задачи можно разделить на интерактивные и неинтерактивные. В неинтерактивных задачах процесс выполнения задачи определен заранее, и пользователь не может на него повлиять. Примерами таких задач являются автоматическое транскрибирование текстов (скажем, судебных заседаний) и автоматическое индексирование мультимедийных данных (радио или телевизионных новостей). Напротив, в интерактивных задачах пользователь сам определяет процесс выполнения задачи, т.е. пользователь ожидает выполнения некоторого действия от компьютера после ввода информации. Примерами интерактивных задач являются управление роботом, интерактивное телевидение, справочные системы.

Современные интерактивные системы обеспечивают взаимодействие между людьми (перевод с одного языка на другой, средства телеконференций, средства поддержки совместной работы) и человека с компьютером. Существует много целей, которые преследует пользовать, взаимодействуя с компью-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

тером: развлечение, получение информации, управление и контроль чем-либо, создание и манипулирование данными, и т.д.

Примерами задач развлечения служат новые интерактивные игры, анимация искусственных персонажей (в мультфильмах), а также интерактивное телевидение. В задачах управления и контроля пользователь совершает некоторые действия для управления определенным процессом. Для этого он может вводить в компьютер некоторые слова или целые фразы. В качестве примера можно привести голосовое управление роботом или использование голосовых команд вместо работы с системой меню или кнопками.

Рис. 1. Классификация многомодальных систем по типам задач.

Новейшие системы безопасности контролируют доступ в здание, используя множество каналов. В запросно-ответных задачах пользователь ведет речевой диалог с системой для получения нужной ему информации. Сейчас активно развиваются голосовые call-центры и справочные системы с использованием систем автоматического распознавания речи. Другим приложением в этой категории являются сервисы, помогающие осуществить заказ различных транспортных услуг (билетов на самолет, поезд или прокат автомобиля), резервирование номера в гостинице, или навигация по городу.

В задачах ввода и манипулирования данными, пользователь создает и управляет данными, которые хранятся в удобной для компьютера форме. В зависимости от сложности данных, эти задачи можно разбить на два типа. Ввод простых данных заключается во вводе отдельных слов, цифр или коротких фраз. К этим задачам относится заполнение форм, адресных книг и т.д. Ввод текстовых и мультимедийных данных гораздо разнообразнее и сложнее. К этой категории относятся системы диктовки, стенографирования или программирования, а также средства разработки пользовательских интерфейсов.

Кроме того, примерами задач, где многомодальные интерфейсы также активно используются, являются «умные» комнаты, системы для образования,

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

мобильные устройства, медицинские системы, системы виртуальной реальности для тренинга и обучения, системы идентификации личности для целей безопасности, web-ориентированные системах и т.д. [12].

Растущему интересу к разработке многомодальных интерфейсов способствует, в значительной степени, идея поддержки прозрачных, гибких и эффективных средств человеко-машинного взаимодействия. От многомодальных интерфейсов ожидают простоты в их изучении и использовании, они более предпочтительны для пользователей во многих приложениях. Многомодальность расширяет спектр пользователей системы и потенциально обеспечивает адаптивность к специфическим условиям функционирования. Такие системы более робастны и устойчивы в работе, чем одномодальные (например, чисто речевые) системы.

Появление многомодальных интерфейсов, основанных на распознавании речи, взгляда, жестов и других выражений естественного поведения — это только начало прогресса к компьютерным интерфейсам, способным функционировать подобно человеческим органам восприятия мира. Такие интерфейсы в будущем смогут непрерывно интерпретировать поступающую информацию от различных визуальных, слуховых и тактильных каналов, которые используются человеком в повседневной деятельности. Одна и та же система сможет отслеживать и объединять информацию от различных датчиков пользовательского интерфейса и окружающего физического пространства для интеллектуальной адаптации к пользователю, текущей задаче и окружению. Будущие адаптивные многомодальные интерфейсы должны реализовывать максимальную функциональность, чтобы получить недостижимую в настоящее время надежность в работе для создания гибких, многофункциональных и персонализированных мобильных систем. Для того чтобы более ясно представлять архитектуру и функциональные возможность многомодальных систем в следующем разделе рассмотрим подробнее входные и выходные модальности, на основе которых осуществляется взаимодействие.

3. Входные и выходные модальности

Очевидно, что многомодальные интерфейсы могут очень сильно различаться по своей внутренней структуре и функциональным возможностям. Прежде всего, это зависит от тех видов модальностей, которые используются в приложении. Модальности в человеко-машинном взаимодействии подразделяются на входные, в которых информация поступает от человека к компьютеру, и выходные, когда потоки информации идут от компьютера к человеку. Как правило, в одной системе объединяются как входные, так и выходные модальности. Обычно в зависимости от того, какие модальности объединяет система, и проводится ее классификация. В табл. 1 представлены входные модальности.

Среди наиболее распространенных классов входных модальностей можно отметить следующие. *Речевой* ввод более предпочтителен, чем традиционные входные модальности в задачах, где заняты руки и глаза (например, при управлении автомобилем), в мобильных приложениях или же там, где речь более удобна (автоматизированные телефонные сервисы). Однако, речь менее удобна в графических задачах (навигация, рисование). *Жестовый* ввод более предпочтителен для указания на графические объекты. *Рукописный* ввод может быть наиболее эффективен для ввода численных данных, а также для заполнения форм и создания пометок.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Таблица 1

Входные модальности

Речь	Жесты	Рукописный ввод	Выражение лица	Традиционный ввод
Слитная Прерывистая Изолированная По буквам	Указание 2D жесты 3D жесты	Слитное написание Печатный текст Отдельные цифры Символы	Движение глаз Направление взгляда Движение губ	Клавиатура Манипулятор-мышь Джойстик Руль Трекбол

Речь — это совместный продукт работы нескольких групп органов: голосовых связок, гортани, легких, движений губ и языка. При ее генерации используются человеческие биомеханические команды для контролирования органов и движения мускулов. С речью связаны как аудио, так и визуальные каналы человека. Уши слышат звук, в то время как глаза видят движения, лица, языка и губ. Кроме того, тактильная информация также используется в речевой среде. Например, слепые люди используют чувство осязания для понимания написанной фразы по методу Брайля.

Многие исследования показывают, что визуальные сигналы важны для лучшего понимания произносимой речи. Например, акцент (привлечение внимания) в речи может быть усилен одним из следующих сигналов: частотой основного тона, поднятием бровей, движением головы, жестом или же комбинацией этих сигналов. Сигналы от визуальных и аудио каналов дополняют друг друга. Это помогает во многих сложных ситуациях при восприятии речи. Некоторые фонемы очень легко спутать на слух (например «м» и «н»), но легко отличить визуально («м» произносится с закрытым ртом, а «н» с открытым). Глядя в лицо собеседнику, нам легче понимать его речь. Слабослышащие люди опираются, в основном, на визуальную информацию, а не на звуковую. Также и системы автоматического распознавания речи, использующие аудиовизуальную информацию работают лучше, чем системы, использующие только аудио информацию.

Первые системы распознавания речи использовали только аудио информацию. Поэтому фоновый шум, чмоканье языком, и иные звуковые артефакты ухудшали качество распознавания. Системы распознавания речи хорошо работали в лабораторных условиях, но в реальных условиях функционирования их качество заметно снижалось. Низкокачественный микрофон или плохой канал передачи звука также ухудшают качество речевого сигнала, а соответственно и точность распознавания.

В табл. 2 представлены неречевые входные модальности, которые используются для человеко-машинного взаимодействия [7, 9]. Распознавание речи не ограничивается только распознаванием того, *что* было сказано, кроме этого учитывается, *как* это было сказано.

Качество голоса и интонация в речи также являются важными характеристиками речи. Параметры голоса всегда индивидуальны для речи. Частота основного тона, громкость, тембр, темп — все это примеры параметров качества голоса. Последовательность акцентов основного тона определяет интонационный контур высказывания. Интонация играет важную роль в речи. Она может передавать информацию о синтаксической структуре и семантике высказывания, а также позволяет определить возраст диктора и его эмоциональное состояние.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Таблица 2

Характеристики неречевых входных модальностей

Входная модальность	Описание и характеристики
Указание и жесты	Форма рук, положение и ориентация рук, а также их движения являются элементами языка жестов, поэтому жестовый ввод можно разделить на 3 категории: указание, двухмерные жесты, трехмерные жесты. При указании используется специальная указка, палец руки или световое перо и активный экран (чувствительный к нажатию). Двухмерные жесты (графические пометки или просто жесты) являются движениями на плоскости, например пометки, нарисованные световым пером на активной панели. Трехмерные жесты являются результатом движения пальцев, рук или головы в трехмерном пространстве. Жесты можно автоматически обрабатывать на компьютере, получив визуальную информацию с видеокамеры, со специальных перчаток или иных устройств, которые могут отслеживать положение. Движения тела и положение всех других частей тела по отношению друг к другу также являются дополнительными источниками информации и используются для интерпретации жестов. Этот язык имеет собственную грамматику и лингвистическую структуру. Комбинация выражения лица, направления взгляда и движения тела также играют важную роль. Например, определенное выражение лица используется для отрицания лексического элемента. Глаза и движения головы наоборот выражают согласие. Кивок головы может быть сигналом для подчеркивания актуальности глагольной конструкции, а также сигналом к принятию решения или указания, что необходимо поставить скобки в предложении. Выражения лица, означающие удивление, гнев, или радость дополняют знаки руками.
Тактильный канал	Исследования показывают, что слепоглухонемые люди могут понимать речь через тактильный процесс, например по методу Тадома. Этот метод основывается на тактильном восприятии артикуляторных элементов речи. Слушающий помещает свои руки на лице и шее говорящего, чтобы чувствовать движения лица, связанные с произнесением речи. Метод Брайля является альтернативным методом для восприятия речи путем тактильных ощущений, но здесь пальцы сканируют специальные таблички для получения информации. Вариации давления на пальцы также могут использоваться для тактильного восприятия речи. Здесь форма давления на каждый палец передает определенную символьную информацию.
Символы и рукописный ввод	Системы распознавания рукописного ввода текста или отдельных символов можно разделить на две категории: автономные системы, которые распознают текст уже написанный (например, распознавание отсканированного текста) и системы реального времени, которые отслеживают текущее положение пера при написании текста на некотором активном экране. Также выделяют различные стили написания: печатными буквами или слитным написанием. Слитное написание текста намного сложнее распознавать, так как тут возникает проблема сегментации на отдельные слова и буквы.
Направление взгляда	Есть ряд прикладных задач, где эта информация может быть полезна, в частности точка фиксации взгляда может рассматриваться как альтернативное указание на объект.
Движения губ	В процессе речеобразования человек естественно использует губы. Визуальная информация может компенсировать недостаток аудио информации в условиях окружающего шума. Глухие же люди полагаются только на эту информацию. Форма губ, позиция языка и видимость зубов позволяют различать элементарные единицы визуальной речи (виземы).
Язык глухонемых	Это пример комбинации нескольких модальностей: жестов и движения губ. При этом форма рук и положение рук используются следующим образом. Одна рука располагается близко к губам и изменяет свою форму синхронно с произносимой речью. Форма руки определяет согласные звуки, в то время как положение руки служит для обозначения гласных звуков. Комбинация губ, формы рук и положения рук служит для уникального представления каждой фонемы. В отличие от языка жестов, основанного на словах, этот язык основывается только на фонемах.
Клавиатура и мышка	Эти модальности сейчас наиболее распространены для взаимодействия человека с компьютером. Клавиатура может состоять всего из нескольких кнопок (пульт управления телевизором) или содержать до сотен кнопок (компьютерные клавиатуры). Мышка используется для отслеживания некоторой траектории движения. К этому же классу относятся и такие устройства как джойстик, трекбол, руль и т.д.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Кроме того, человек при восприятии речи еще использует и другие модальности — дополнительные визуальные и тактильные способы получения информации, ведь человеческие чувства включают в себя: слух, зрение, вкус, осязание, обоняние. Современные многомодальные компьютерные системы могут частично моделировать слух, зрение и осязание.

В табл. 2 были представлены основные виды входных модальностей, которые могут быть использованы в многомодальных системах. Теперь перейдем к рассмотрению выходных модальностей, которые служат для информирования пользователя о событиях происходящих вокруг него и внутри самой системы. В табл. 3 представлены выходные модальности, их основные характеристики, принципы воспроизведения и примеры использования.

Таблица 3

Характеристики выходных модальностей

Выходная модальность	Описание и характеристики
Текст	Когда необходимо получить большой объем важной информации, то более эффективно ее прочитать в виде текста, чем прослушать. Причина этого в том, что речь сохраняется у слушающего в кратковременной памяти головного мозга, в том время как прочитанный текст в визуальной области долговременной памяти.
Изображения	Картинки, фильмы или анимация служат примерами иллюстрирования текста или речи. В некоторых приложениях показ визуальной информации более эффективен, чем прослушивание той же информации в виде речи.
Звук	Звук присутствует во всех кинофильмах, музыке и сопровождает речь и визуальную информацию. В человеко-машинных интерфейсах звук также может использоваться в виде гудков или специальных сигналов, привлекающих внимание пользователя.
Синтез речи	При построении таких систем синтеза один из подходов состоит в хранении акустических образов записанных участков речи (например, целых слов или же фраз) и последующего их проигрывания. Второй подход состоит в том, чтобы разделить текст на фонемы и вычислять различные параметры фонем (основная частота, форманты, ударение в слове), а затем при синтезе речь генерируется путем подбора и конкатенации наиболее подходящих моделей фонем.
Голосовой тракт	Синтез речи, снабженный артикуляцией должен моделировать аппарат голосового тракта человека. Некоторые системы синтеза речи основаны на связи между артикуляторными жестами для производства речи и акустическим выводом речи. Проблема здесь состоит в том, чтобы найти правильное отображение между акустическими параметрами и геометрическими (центр языка, толщину губ, угол поворота челюсти) параметрами представления речевого тракта. Вычисление параметров голосового тракта может выполняться, например, посредством рентгеновских снимков.
Оптическая генерация	Как уже было сказано, визуальная речь улучшает понимание речи у слабослышащих людей. Поэтому говорящая голова, которая может двигать губами синхронно с синтезом речи, значительно повышает качество восприятия такой речи любым человеком. Здесь также исследуется проблема выражений лица, связанных с эмоциями и интонацией.
Механическая генерация	Были разработаны системы, которые используют механические лица для моделирования системы Тадома для слепоглухонемых людей. Они симулировали движения губ совместно с движениями челюстей. Примером такой системы для слепых является система MEDITOR [2], в которой применяется синтез речи и дисплей Брайля.
Тактильный вывод	Этот тип вывода реализуется при помощи специальных устройств, например, перчаток или шлемов виртуальной реальности.

Значительный интерес уделяется многомодальным интерфейсам, которые объединяют различную визуальную информацию, такую как направление взгляда человека, выражение лица, жесты руками или частями тела. Эти технологии незаметно (пассивно) и постоянно отслеживают поведение пользователя и не требуют подачи конкретных команд компьютеру. В отличие от этих модальностей, речевой ввод и световое перо используются для выдачи опре-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

деленных команд указывающих намерение пользователя, поэтому такие модальности являются активными в ходе диалога.

Также важным направлением исследований является разработка смешанных интерфейсов, которые объединяют как пассивные, так и активные модальности в одной системе. Такие интерфейсы, как правило, являются каскадированными по времени. Это позволяет определить стратегии оптимальной обработки многомодальной информации, т.е. возможно использование информации от первой модальности (например, направления взгляда), чтобы ограничить интерпретацию следующих видов ввода (например, жестового или речевого). Смешанные интерфейсы потенциально могут обеспечивать большую функциональность и робастность.

В процессе проектирования многомодальных систем разработчики апеллируют к поведению человека в данной ситуации или общению между людьми. Что неудивительно, поскольку при построении любой системы искусственного интеллекта, естественно, что за основу берется прототип — человек. Кроме того, система, построенная по привычным человеку принципам взаимодействия, будет удобна пользователю в работе с ней. Поэтому в следующем разделе будут рассмотрены когнитивные аспекты многомодальных интерфейсов и некоторые вопросы, связанные с познавательными и поведенческими способностями человека.

4. Когнитивная наука и многомодальные интерфейсы

Создание новых многомодальных архитектур и систем обуславливается двумя предпосылками. Во-первых, когнитивная наука, изучающая человеческие механизмы восприятия и межчеловеческое взаимодействие, обеспечила фундаментальную информацию для моделирования поведения пользователя, а также информацию о том, как должны быть построены системы распознавания и организованы многомодальные архитектуры. В частности, когнитивная наука предоставляет необходимые знания о моделях естественной интеграции информации, которые служат для объединения движений губ и мимики лица с речевым вводом, а также знания о том, как человек использует жесты руками в различных речевых диалогах.

В сложной природе многомодального взаимодействия когнитивная наука приобретает особую важность при разработке робастных и эффективных многомодальных систем. Реалистичные автоматические симуляторы поведения также играют критическую роль в создании новых прототипов систем. На стадии планирования систем, разработки дизайна и первичного тестирования системы такие модели могут использоваться для планирования человеко-машинного взаимодействия. Так на первом этапе дизайна системы возможна подмена компьютерной системы человеком-разработчиком и общение с пользователем, который полагает, что работает с реальной полнофункциональной компьютерной системой (рис. 2). В ходе таких экспериментов разработчик сам эмулирует работу системы, контролирует поведение пользователя, отслеживает многомодальный ввод и выдает ответы системы настолько точно насколько возможно [14]. Этот процесс позволяет выстроить эффективные модели диалогов в будущей многомодальной системе и учесть возможное поведение пользователей. Такой подход получил название Волшебник страны Оз (Wizard of Oz) [4].

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

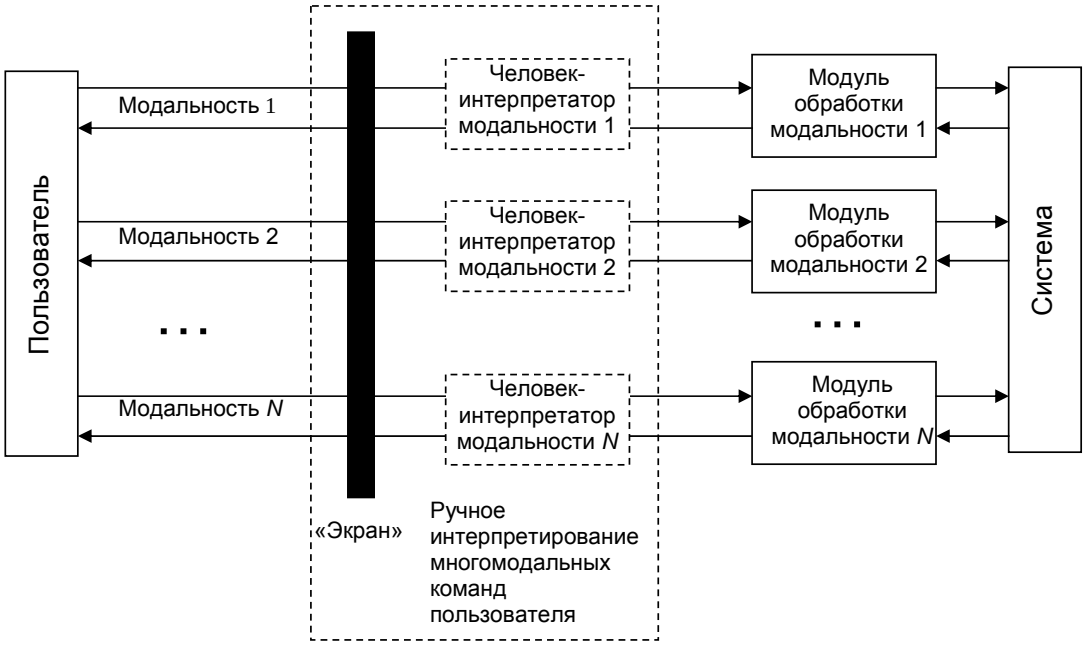

Рис. 2. Моделирование работы многомодальной системы в режиме Wizard of Oz.

Для подобного моделирования разрабатывается специальное программное обеспечение, максимально приближенное к симулированию действий разрабатываемой многомодальной системе.

Возможности разработки многомодальных систем во многом зависят от знания методов естественной интеграции, которые объединяют различные входные модальности. В частности, дизайн новых систем зависит от знания свойств различных модальностей и их информационного контекста, характеристик многомодального языка, а также интеграции и синхронизации многомодального взаимодействия с пользователем. Также очень важно представлять, когда пользователи взаимодействуют с системой многомодально, а когда нет.

Во время традиционной межчеловеческой коммуникации люди почти всегда взаимодействуют многомодально, поэтому неудивительно, что пользователи хотят использовать многомодальный диалог в очень широком спектре приложений. Например, около 95% пользователей предпочитают взаимодействовать многомодально с интерактивными картами, свободно используя как речевой ввод, так и указание пальцем или световое перо. При этом речевой ввод служит для описания объектов и событий, множеств объектов, объекты вне поля зрения, а также для подачи команд для выполнения действий. Рукописный или перьевой ввод предпочтителен при работе с цифрами, символами, графическим контекстом, а особенно, для получения положения и формы пространственно-ориентированной информации, расположенной на графическом дисплее или карте.

Также большинство пользователей используют речь и жесты, при манипулировании графическими объектами на экране (например, для создания, перемещения, удаления объектов). Считается, что рост производительности — это основное преимущество многомодальных систем, поскольку они способны обрабатывать входные модальности параллельно. Это верно, и достаточно часто они улучшают эффективность взаимодействия, особенно при манипулировании графической информацией [10]. В ходе исследований было выяснено, что по сравнению с чисто речевыми интерфейсами многомодальные интерфейсы (ре-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

чевые и указательные) обеспечивают прирост 10% по скорости выполнения визуально-пространственной задачи, однако не наблюдается значительного прироста эффективности в «разговорных» или количественных задачах. Кроме того, увеличивается эффективность при комбинировании речевой модальности с жестами для манипулирования трехмерными объектами, по сравнению с одномодальным вводом. Например, система QuickSet, использующая речевой и перьевой ввод оказалась в 4 раза эффективнее, чем традиционный интерфейс на основе мыши и клавиатуры [6]. При тестировании также учитывалось время, необходимое для коррекции ошибок распознавания.

В то же время, очевидно, что число источников информации или модальностей, которые доступны человеку или пользователю, ограничено. Также и собеседник, и многомодальные системы ограничены по числу и типу входных модальностей, которые они могут использовать (распознавать/воспринимать) (рис. 3.). Кроме того, пользователи могут комбинировать виды ввода во время человеко-машинного взаимодействия, или общаться многомодально, или использовать только одну модальность. Хотя пользователи и предпочитают взаимодействовать многомодально, это абсолютно не гарантирует, что они будут подавать каждую команду системе многомодально.

Рис.3. Многомодальное взаимодействие между людьми и с компьютером.

При разработке системы необходимо решить: пользователи будут работать одномодально или многомодально? В системе, где предусмотрен речевой и перьевой ввод пользователи обычно смешивают одномодальное и многомодальное взаимодействие [10]. Исследования показывают, что при работе с визуально-пространственными объектами, пользователи примерно в 20% случаев управляют многомодально, а остальное время отдаются только речевые команды или символы, сделанные световым пером. Предсказать, будет ли пользователь выражать команды многомодально можно, зная тип действия, которое пользователь выполняет в данный момент. В частности, пользователи почти всегда выражают команды многомодально, когда описывают пространственную информацию о расположении, размере, ориентации, форме объекта. Почти всегда пользователи подают команды многомодально, когда им нужно добавить, передвинуть, модифицировать некоторые объекты на карте или вычислить расстояние между объектами. Также многомодальные команды, как правило, используются при выделении элементов из большого массива, например,

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

городов на карте. Однако при выполнении общих действий без пространственного компонента (например, распечатка карты) пользователи совершенно не используют многомодальных команд, а только лишь речевые высказывания. Это все говорит о том, что грамотно построенная многомодальная система должна уметь распознавать случаи, когда пользователи не общаются многомодально, т.е. системе нужны знания о типе действий, которые используются в приложении (в частности, будет ли манипулирование пространственной информацией) для принятия решения о том, нужно ли применять многомодальный интерфейс.

В интерфейсе, который обрабатывает пассивные или смешанные виды ввода, всегда есть, по крайней мере, один пассивно отслеживаемый источник ввода непрерывной информации (взгляд или позиция головы). В этих случаях все команды пользователя будут многомодальными и основной проблемой при работе таких систем становится сегментация и интерпретация каждого из непрерывных потоков данных для выполнения определенных действий в приложении. В случае смешанного интерфейса (например, направление взгляда и ввод при помощи мышки) уместно выявлять активные формы пользовательского ввода, которые могут быть точно и быстро обработаны как многомодальные события.

Первые многомодальные системы фокусировались на простом выделении объектов или определении положения на дисплее, а не на рассмотрении широкого спектра методов многомодальной интеграции. В системе “Put That There” 1980 года использовались речь и указания [5]. В этой системе семантическая обработка базировалась на речевом вводе, а ключевое слово «здесь» превращалось в координаты точки на дисплее. В более поздних системах вместо указания рукой также применялась информация о точке фиксации взгляда пользователя.

К сожалению, концепция такого многомодального взаимодействия как речь и указание создает ограничения на добавление новых модальностей для выделения объектов (как это делается посредством мышки). В противоположность этому, модальности, которые передают рукописный ввод, жесты руками и выражения лица способны генерировать символьную информацию, причем с гораздо большей семантической нагрузкой, чем простое указание или выделение. Лингвистический анализ спонтанных жестов руками показывает, что во время межличностной многомодальной коммуникации простые указания (рукой или пальцем) составляют менее 20% от всех жестов, производимых человеком. Данные когнитивной науки и моделирования поведения пользователя показывают, что система, использующая только речевой ввод и указательные жесты, неспособна обеспечить пользователей полным набором функциональных возможностей. Ясно, что для будущих многомодальных систем нужен более широкий набор методов для интеграции многомодальной информации.

В многомодальных речевых и рукописных интерфейсах пользователи могут выбирать тот вид ввода информации, который в конкретной ситуации обеспечивает меньшее количество ошибок. Они могут предпочесть более быстрый речевой ввод и переключаться на рукописный ввод для набора иностранных фамилий или длинных последовательностей цифр. Язык взаимодействия часто упрощается при многомодальном общении, что может значительно уменьшить сложность обработки естественного языка и соответственно уменьшить количество ошибок распознавания. Кроме того, при реальной работе пользователи

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

предпочитают переходить от одного вида ввода к другому после возникновения ошибок распознавания.

Для корректировки ошибок системы существуют хорошо продуманные многомодальные архитектуры с двумя семантически насыщенными видами ввода, которые могут поддерживать взаимное устранение неоднозначностей входных сигналов [8]. Например, если пользователь военной системы говорит «окопы», а распознаватель речи выдает «окоп», то независимое параллельное распознавание нескольких графических объектов на карте, обведенных пользователем при помощи светового пера, может в результате привести к правильной совместной интерпретации. Это устранение неоднозначности может происходить в многомодальной архитектуре, даже когда распознаватель речи оценивает интерпретацию входной фразы «окопы» как не самую лучшую и вероятную гипотезу.

Таким образом, в многомодальной архитектуре взаимное устранение неоднозначностей обеспечивает исправление ошибок на отдельных одномодальных уровнях распознавания и ведет к более стабильной и робастной работе системы. Оптимальное качество распознавания достигается за счет включения дополнительных видов ввода, обеспечивающих дублирование входной информации по разным каналам человеко-машинного взаимодействия. Результаты исследований показывают, что хорошо продуманная многомодальная система может функционировать более робастно, чем одномодальная система для широкого спектра реальных пользователей и используемых контекстов и задач [8]. Например, во время аудио-визуального восприятия речи и движений губ наблюдается улучшение качества распознавания речи, как для носителей языка, так и пользователей с акцентом, за счет использования технологии устранения неоднозначностей.

Иногда считают, что сигналы от разных информационных потоков синхронизированы во времени. Это предположение означает, что временное перекрытие поступающих сигналов определяет, какие сигналы необходимо объединять в одном многомодальном акте коммуникации во время обработки. В этом случае системы с речевым и жестовым вводом могут успешно обрабатывать такие выражения, как «удалить <этот> квадрат» или «нарисовать круг <здесь>». При этом слово «этот» говорится, чтобы уточнить, какой именно объект должен быть удален. Однако многие эксперименты доказывают, что пользователи редко говорят такими четкими терминами или же говорят не одновременно с указанием, и при этом речевой сигнал не перекрывается во времени с указанием на объект. Более того, было установлено, что менее 25% пользовательских команд содержат речевую информацию, которая перекрывается по времени с указанием [10].

Во время взаимодействия с компьютером посредством речевого и перьевого ввода, как правило, перьевой ввод опережает голосовой, и интервал между этими сигналами может достигать 1–2 секунды. Это согласуется с лингвистическими данными когнитивной науки [9]. Спонтанные жесты обычного человека и жесты языка глухих несколько опережают само речевое высказывание во время межчеловеческой коммуникации. Причем интересно, что степень отставания речи от жестов больше в тематико-ориентированных языках (например, китайском), чем в предметно-ориентированных (английском, испанском или русском). Более того, в системах аудио-визуального распознавания речи, использующих совместно аудиоинформацию и видеоинформацию о форме губ отмечается, что эти модальности не являются полностью синхронизированны-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ми [15]. Иными словами, хотя два вида ввода могут быть независимыми и синхронизированными, но эта синхронность не означает полной одновременности. Поэтому при разработке многомодальных систем не всегда можно рассчитывать на то, что сигналы будут перекрывающимися во времени, и это необходимо учитывать в многомодальной архитектуре.

Сейчас ведется разработка методов интеграции асинхронных модальностей и временного каскадирования, возникающего в трех или более модальностях, таких как направление взгляда, жесты и речь. Нужно отметить, что при создании новых многомодальных архитектур должны использоваться данные о порядке обработки модальностей и приблизительном интервале времени между сигналами для определения вероятности того, что высказывание является многомодальным (или одномодальным), чтобы установить временные пороги в методе многомодального объединения.

Среди пользователей существуют индивидуальные и групповые особенности в предпочтении различных видов коммуникации, и многомодальные интерфейсы позволяют различным группам пользователей выбирать подходящую форму общения и контролировать свое взаимодействие с компьютером. Соответственно, они потенциально могут объединить более широкий спектр пользователей, чем традиционные интерфейсы, включая пользователей разных возрастов, уровней способности, национальных языков, пользователей со специфическими потребностями, с ухудшенным восприятием и другими временными или постоянными недостатками (например, инвалидов) [13]. Например, плохо видящий пользователь предпочтет речевой ввод на входе и синтез речи на выходе системы. Напротив, плохо слышащий пользователь или пользователь с акцентом предпочтет тактильный, жестовый или письменный ввод информации. Также переключение между каналами ввода информации может быть эффективным в случае повреждения или отказа одного из них, особенно в случае длительного периода функционирования.

В то же время индивидуальные отличия следует учитывать при интеграции модальностей и это вносит дополнительные сложности при проектировании многомодальных систем. Так, эксперименты показывают, что в системах речь/перо можно выделить две категории пользователей: (1) те, кто совмещает речевые сигналы с символами от светового пера во времени, т.е. выдают многомодальные информацию одновременно и синхронно; (2) те, кто вводят данные последовательно (сначала пером, а затем голосом), т.е. синхронизируют сигналы последовательно, причем речь может запаздывать даже на несколько секунд. Эти различные методы интеграции могут быть идентифицированы в начале взаимодействия пользователя с системой, а затем использованы во время дальнейшего взаимодействия.

Метод интеграции для каждого пользователя определяется заранее и остается постоянным в процессе взаимодействия, несмотря на то, что методы интеграции отличаются для различных пользователей. Как было уже сказано, есть значительный разброс во времени, на которое ручные жесты предшествуют речи для различных лингвистических групп (китайцев, американцев, русских). Это означает, что многомодальные архитектуры должны быть способны адаптироваться к временным порогам для различных групп пользователей, что потенциально должно приводить к большей точности распознавания и скорости взаимодействия.

Существуют как индивидуальные, так и культурные различия между пользователями в методах интеграции модальностей. Можно определить степень, в

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

которой расходятся по времени речь и движения губ в шумном окружении или для различных народов. Например, движения губ во время речеобразования более синхронизированы по времени для японских дикторов, чем для американских [9]. Наконец, неносители языка, глухонемые дикторы, а также пожилые люди больше опираются на информацию от движений губ, чем на слуховую информацию. Кроме того, пол, возраст и другие индивидуальные различия являются важными для распознавания взгляда, а также интеграции речи и взгляда. Также, необходимо знать и учитывать при разработке, что в разных этносах и культурах жесты могут иметь различное значение, иногда даже совершенно противоположное [9].

Считается, что содержание (смысл действия), передаваемое различными модальностями во время многомодальной коммуникации содержит большую избыточность, т.е. дублируется по разным информационным каналам. Однако, в ходе исследований было выяснено, что наиболее часто проявляется «дополнительность» содержимого многомодального пользовательского ввода. Например, речевой и перьевой ввод последовательно вносят вклад в общую семантическую информацию многомодального коммуникативного акта: при помощи речи определяется предмет, глагол (тип действия), и субъект, а пространственная информация и местоположение обычно передаются при помощи пера. В действительности, главная «дополнительность» между речью и перьевым вводом содержится в визуально-пространственном семантическом содержании, и это причина того, что такие модальности являются хорошей комбинацией для визуально-пространственных приложений. Любая пространственная информация четко определяется перьевым вводом, а речевое описание наиболее подходит для временной и непространственной информации. Даже при многомодальной коррекции ошибок распознавания системы, когда пользователи особенно заинтересованы в точном распознавании их информации, эта комбинация модальностей передает информацию избыточно менее чем в 1% случаев [8].

Лингвистами также замечено, что во время межличностной коммуникации информация от спонтанной речи и жестов руками, в основном, является дополняющей, а не дублирующей. Можно сделать вывод, что имеющиеся данные отражают важность «дополнительности» как главного организационного направления многомодальной коммуникации. Поэтому разработчикам многомодальных интерфейсов следующего поколения не следует полагаться на дублирование информации от модальностей. Преимущество использования дополнительной информации заключается в том, что результирующая многомодальная архитектура может функционировать более робастно, чем одномодальная или многомодальная, но без дополняющих модальностей.

Коммуникативные каналы могут иметь огромное влияние на форму языка, передаваемого по ним. Многие исследования показывают, что существуют лингвистические особенности многомодального языка, и этот язык качественно отличается от разговорного или формального текстового языка. В основном характеристики отличаются в таких параметрах как лаконичность, семантическое содержание, синтаксическая сложность, порядок слов, степень неопределенности и т.д. Во многих случаях многомодальный язык лингвистически проще, чем разговорный. В частности, сравнение показывает, что тот же самый пользователь при многомодальном общении с интерактивной картой использует меньше слов, составляет более короткие предложения, и менее сложные пространственные описания, чем при чисто речевом общении. Следующий пример пока-

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

зывает типичное речевое высказывание пользователя для открытия области на карте: «Покажи район от улицы Гагарина до проспекта Космонавтов». Но тот же самый пользователь, выполняющий эту же задачу, может очертить определенную область на карте специальным пером и сказать: «Покажи <этот> район». Это объясняется тем, что людям тяжело объяснить пространственную информацию при помощи речи и лучше это сделать перьевым вводом.

При данном виде коммуникации лингвистическая неопределенность, которая типична для речевого языка, заменяется более определенными командами. Например, пользователь делает некоторый запрос для вычисления расстояния на карте, используя речевой ввод: «Какое расстояние между Исаакиевским собором и Мариинским театром?». Когда же этот запрос делается многомодально, то тот же самый пользователь обводит кружком собор, а затем театр и выдает точную команду: «Какое расстояние от <сюда> до <сюда>?». Очевидно, что фразы в многомодальном языке общения становятся короче по сравнению с одномодальным языком. Но, хотя многомодальный язык и отличается от разговорного языка, он не обязательно проще. Например, многомодальный язык (английский) отличается от канонического порядка слов в предложении (субъект-глагол-объект-местоположение), который свойственен речи, а также от формальных текстовых языков. Исследования показывают, что 95% слов, определяющих местоположение, стоят на первом месте в предложении при многомодальном взаимодействии. Все это требует того, чтобы многомодальные базы данных, статистические модели языка, и алгоритмы обработки естественного языка, определяющие многомодальный язык, были продуманы до начала разработки любой многомодальной системы.

В заключение раздела приведем некоторые размышления, которые ведущие исследователи назвали как «10 мифов» (или заблуждений) в области многомодальных интерфейсов, сформировавшихся у пользователей и разработчиков систем [11]:

1) в многомодальной системе пользователи всегда будут взаимодействовать многомодально;
2) речь — главная модальность в любой многомодальной системе;
3) Речевой ввод и указание — доминирующая комбинация модальностей;
4) многомодальный ввод содержит одновременные и синхронизированные сигналы;
5) многомодальный язык лингвистически не отличается от одномодального языка;
6) избыточность содержимого разных модальностей — основа интеграции в многомодальных архитектурах;
7) объединение технологий распознавания отдельных модальностей ведет к объединению ошибок, в результате система становится менее надежной;
8) асе пользовательские многомодальные команды могут объединяться единым образом;
9) различные модальности способны передавать сопоставимое содержание;
10) повышенная производительность системы — основное преимущество многомодального интерфейса.

Заметим еще раз, что это именно ошибочные принципы и на протяжении всего этого раздела были приведены аргументы по неадекватности каждого из них. Поэтому при проектировании многомодальных интерфейсов необходимо

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

очень осторожно подбирать функции каждой входной и выходной модальности, используя привычные способы общения между людьми. Это является наиболее простой проверкой естественности интерфейса, и поэтому при проектировании разработчики довольно часто берут за основу взаимодействие человека с человеком в аналогичной задаче.

5. Заключение

Решением проблем эффективного взаимодействия человека с компьютером в естественной форме сейчас занимаются специалисты разных научных областей. Инженеры, математики, программисты разрабатывают более эргономичные программные и аппаратные средства управления компьютерами. В свою очередь лингвисты, физиологи, психологи и специалисты других специальностей изучают поведение человека, а также механизмы восприятия, обработки информации для того, чтобы понять основу процессов мышления. Междисциплинарный подход к разрешению проблем построения эффективных и естественных интерфейсов кажется наиболее перспективным на сегодняшний день.

Многомодальный характер коммуникации наиболее приближен к способу общения между людьми. В диалоге человек использует те модальности, которые доступны, удобны, понятны ему и собеседнику. В этом преимущество многомодальных интерфейсов, поскольку пользователь может выбирать модальности, которые естественны для него и ведут к минимуму ошибок распознавания в системе. Кроме того, за счет передачи одинаковой семантической информации разными модальностями повышается робастность системы. Многомодальные интерфейсы особенно зарекомендовали себя в картографических системах, умных комнатах, а также различных приложениях для инвалидов и людей с ограниченными возможностями. Выбор доступных для них модальностей позволяет оперировать со сложной техникой, несмотря на физические ограничения. Многомодальные интерфейсы только начинают развиваться, и их успех во многом зависит от полноты знаний о способах и механизмах взаимодействия между людьми.

Литература

1. *Карпов А. А., Ронжин А. Л.* Многомодальные интерфейсы в автоматизированных системах управления // Известия вузов. Приборостроение. 2005. Вып. 48. С. 9–14.
2. *Bellik Y.* MEDITOR: a Multimodal Text Editor for Blind Users // Proc. of ACM UIST'96, Ninth Annual Symposium on User Interface Software, Seattle, Washington, USA, 1996.
3. *Benoit C.., Martin J. C., Pelachaud C., Schomaker L., and Suhm B.* Audiovisual and Multimodal Speech Systems // Handbook of Standards and Resources for Spoken Language Systems / D. Gibbon (Ed.). Dordrecht: Kluwer Academic Publishers, 2000. 544 p.
4. *Bernsen N. O., Dybkjær H. and Dybkjær L.* Designing Interactive Speech Systems. From First Ideas to User Testing. New York: Springer Verlag, 1998. 276 p.
5. *Bolt R. A*.. Put-That-There: Voice and Gesture at the Graphics Interface // Computer Graphics. 1980. Vol. 14, no. 3. P. 262–270.
6. *Cohen P. R., Johnston M., McGee D., Oviatt S.* Quickset: Multimodal Interaction for Distributed Applications // Proc. of the Fifth ACM International Multimedia Conference. New York: ACM Press, 1997. P. 31–40.
7. *Liddel S. K.* Structures for Representing Handshape and Local Movement at the Phonemic Level // Theoretical Issues in Sign Language Research / *Fischer S. D.* (Ed.). University of Chicago Press. 1990. P. 37–65.

8. *Oviatt S L.* Mutual Disambiguation of Recognition Errors in a Multimodal Architecture // Proceedings of the Conference on Human Factors in Computing Systems (CHI'99). New York: ACM Press, 1999. P. 576–583.
9. *Oviatt S. L*. Multimodal Interfaces // The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, *Jacko J. and Sears A.* (Eds.). Mahwah, NJ: Lawrence Erlbaum Assoc. 2003. P. 286–304.
10. *Oviatt S. L.* Multimodal Interactive Maps: Designing for Human Performance // Human-Computer Interaction. Special Issue on Multimodal Interfaces. 1997. Vol. 12. P. 93–129.
11. *Oviatt S. L.* Ten Myths of Multimodal Interaction // Communications of the ACM. 1999. Vol. 42. P. 74–81.
12. *Ronzhin A. L., Karpov A. A., Timofeev A. V., Litvinov M.V.* Multimodal Human-Computer Interface for Assisting Neurosurgical System // Proc. of 11-th International Conference on Human-Computer Interaction HCII-2005, Las Vegas, Nevada, USA, Mira Digital Publishing, Las Vegas, 2005.
13. *Ronzhin A., Karpov A*. Assistive Multimodal System Based on Speech Recognition and Head Tracking // Proc. of 13-th European Signal Processing Conference (EUSIPCO-2005), Antalya, Turkey, 2005.
14. *Salber D., Coutaz J.* Applying the Wizard of Oz Technique to the Study of Multimodal Systems // Proc. of East/West Human Computer Interaction, Moscow, 1993. P. 219–230
15. *Železný M., Císar P., Krnoul Z., Ronzhin A., Li I., Karpov A.* Design of Russian Audio-Visual Speech Corpus for Bimodal Speech Recognition // Proc. of 10-th International Conference SPECOM'2005, Patras, Greece, 2005. P. 397–400.

Труды СПИИРАН. 2006. Вып. 3. Т. 1. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2006. Issue 3. V. 1. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru