А. А. Карпов, Когнитивные исследования ассистивного многомодального интерфейса для бесконтактного человеко-машинного взаимодействия, *Информ. и её примен.*, 2012, том 6, выпуск 2, 77–86

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2012. Т. 6. Вып. 2. С. 77–86

КОГНИТИВНЫЕ ИССЛЕДОВАНИЯ АССИСТИВНОГО МНОГОМОДАЛЬНОГО ИНТЕРФЕЙСА ДЛЯ БЕСКОНТАКТНОГО ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ*

А. А. Карпов[1]

Аннотация: Представлены результаты исследований многомодального пользовательского интерфейса, предназначенного для бесконтактного управления персональным компьютером при помощи речевого ввода и указательных жестов/движений головой. Данный многомодальный интерфейс использует низкобюджетное аудио- и видеооборудование для одновременного захвата многоканальных сигналов и обеспечивает универсальный доступ к компьютерным системам как обычных операторов для бесконтактной (без использования рук) работы с компьютером, так и пользователей с ограниченными физическими возможностями (с проблемами двигательных функций рук или даже не имеющих рук/пальцев). Описаны методики и результаты количественной оценки производительности бесконтактного человеко-машинного взаимодействия с применением элементов когнитивных экспериментов и сравнение с результатами для стандартных контактных способов указательного ввода информации.

Ключевые слова: многомодальный интерфейс; распознавание речи; машинное зрение; ассистивные информационные технологии

1 Введение

Многие люди не могут полноценно работать с компьютерными системами (печатать тексты, работать в Интернете, рисовать, и т. д.) по причине физических ограничений, например ампутации рук в результате войн, аварий, врожденных дефектов или парализации рук в результате болезней. Для таких людей и создается многомодальный пользовательский интерфейс бесконтактного взаимодействия с компьютером посредством речевого ввода и отслеживания осмысленных движений (жестов) головы или тела человека. Согласно общепринятому определению, «жест» (от лат. *gestus* — движение тела) — это некоторое действие или движение человеческого тела или его части (например, рук, головы или глаз), имеющее определенное значение или смысл. В этом смысле жестом может являться кивок, покачивание или наклон головы, а также указательные жесты, когда пользователь головой показывает на определенное направление движения.

Разработанный за последние годы в лаборатории речевых и многомодальных интерфейсов СПИИРАН ассистивный (предназначенный для помощи) пользовательский интерфейс получил сокращенное название ICanDo («Я могу делать»), что расшифровывается как “Intellectual Computer AssistaNt for Disabled Operators” («Интеллектуальный компьютерный помощник для операторов-инвалидов») [1]. Он снабжен программными технологиями автоматического распознавания русской речи / голосовых команд и технического зрения для отслеживания движений головы (указательных жестов) с целью управления курсором мыши на экране дисплея, что повышает естественность и эффективность человеко-машинного взаимодействия. Речевое взаимодействие является наилучшей альтернативой любым устройствам ввода для задачи набора текста на компьютере как для пользователей-инвалидов, так и для обычных пользователей. Видео- и аудиосигналы одновременно и параллельно захватываются одним аппаратным устройством — цифровой видеокамерой (веб-камерой) и синхронно обрабатываются в многомодальном интерфейсе.

Альтернативой программному человеко-машинному интерфейсу для пользователей без верхних конечностей могут служить различные аппаратные устройства для управления графическим интерфейсом компьютера, например аппаратно-программные устройства слежения — трекеры головы (зарубежная система InterTrax, которая использует гироскоп; система SmartNav, которой необходим инфракрасный приемопередатчик;

* Работа выполнена при поддержке Минобрнауки РФ в рамках ФЦП «Исследования и разработки», госконтракт № 11.519.11.4025; Совета по грантам Президента РФ, проект МК-1880.2012.8; фонда «Научный Потенциал» и Комитета по науке и высшей школе Правительства Санкт-Петербурга.

[1] Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), karpov@iias.spb.su

оптическая система HeadMouse Extreme). Чтобы использовать данные системы для управления курсором мыши на экране дисплея, пользователь должен надеть на голову специальное устройство (шлем или очки виртуальной реальности со встроенным микроминиатюрным гироскопом в случае InterTrax, либо специальную конструкцию со светоотражающими метками в случае SmartNav или HeadMouse Extreme). Кроме того, для этой задачи могут также применяться специальные устройства со светодиодами и аккумуляторами, например комплект для ассистивного управления компьютером КАУ-09-1, разработанный в ЗАО НПК ФАТУМ, или цветными реперными (контрольными) точками-мишенями, которые крепятся на специальном шлеме, надеваемом на голову, например аппаратная система «Шлемомышь» [2], разработанная лабораторией компьютерной графики факультета вычислительной математики и кибернетики МГУ. Реперные точки на таких устройствах отслеживаются посредством инфракрасной либо цифровой видеокамеры. Однако пользователи и психологи говорят о том, что люди не хотят использовать для человеко-машинного взаимодействия специальные, носимые на голове или теле аппаратные устройства, значительно снижающие естественность взаимодействия и мобильность передвижения из-за наличия проводов, кабелей, аккумуляторов для автономной работы, их общей громоздкости и технических сложностей в калибровке и установке. Кроме того, люди без рук не могут надеть такое устройство сами себе на голову.

Для некоторых задач и для определенных категорий пользователей-инвалидов (например, парализованных лежачих людей) перспективно применение аппаратно-программных систем для трекинга направления взгляда. Такие системы в мире существуют, например зарубежные трекеры глаз Eyegaze System или Visual Mouse, но их использование и внедрение на практике осложняется тем, что необходимо использовать очень дорогие высокоскоростные цифровые видеокамеры высокой четкости с большим разрешением, так как область глаза незначительна по размеру и сложна в распознавании. Кроме того, как показывают когнитивные исследования [3], использование отслеживания направления взгляда для управления курсором мыши намного сложнее для обучения и хуже, чем отслеживание движений головы, по следующим показателям: производительность, эмоциональная нагрузка на пользователя, удобство использования, эргономичность.

Кроме того, в качестве альтернатив бесконтактному взаимодействию можно упомянуть управление манипулятором-мышью с использованием ног вместо рук или специальный тактильный манипулятор, функционирующий за счет изменения положения центра масс тела человека, сидящего на специальной «подушке» [4]. В будущем, возможно, будут доступны и системы взаимодействия на основе прямого интерфейса мозг–компьютер, во всяком случае, разработки в области нейроинформатики активно ведутся как за рубежом, так и в России.

В ассистивном интерфейсе ICanDo, которому посвящена данная статья, реализованы и применены программные средства компьютерного зрения для обнаружения лица человека в оптическом потоке на основе характерных органов/черт лица (нос, глаза, губы) без использования искусственных маркеров/мишеней и специализированных, носимых человеком, устройств, что выгодно отличает его от имеющихся аналогов. Программная система взаимодействия не накладывает дополнительных ограничений на пользователя и обеспечивает естественность и комфорт при бесконтактной работе с компьютером. Применяемые в интерфейсе голосовые команды, распознаваемые автоматической системой, являются прекрасной альтернативой стандартным органам ввода информации (клавиатура) как для инвалидов без рук или пальцев рук, так и для обычных пользователей.

2 Архитектура ассистивного многомодального интерфейса пользователя

Ассистивный интерфейс человеко-машинного взаимодействия относится к классу многомодальных пользовательский интерфейсов [5] и использует две естественные входные модальности: речь на русском языке и указательные жесты — движения головы (вверх, вниз, вправо, влево и в любых промежуточных направлениях). Обе модальности являются активными [6] и инициируются напрямую человеком, поэтому они непрерывно отслеживаются и обрабатываются интеллектуальными подсистемами интерфейса. Каждая из модальностей передает свою семантическую информацию: положение головы определяет положение курсора мыши на рабочем столе компьютера в конкретный момент времени, а речевой сигнал передает информацию о действии, которое должно быть выполнено с некоторым объектом графического пользовательского интерфейса. На рис. 1 представлена архитектура аппаратно-программного комплекса ассистивного многомодального интерфейса.

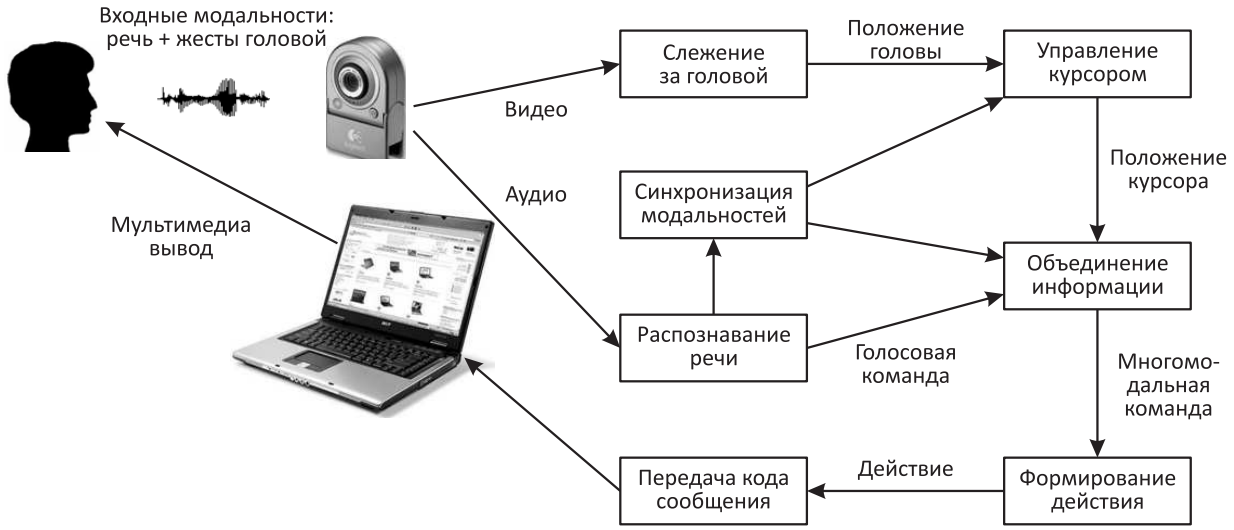


Рис. 1 Архитектура ассистивного многомодального пользовательского интерфейса

Многомодальный интерфейс способен распознавать несколько десятков голосовых команд для управления компьютером (например, «открыть», «сохранить», «левая», «правая», «ввод» и т. д.). Всего система содержит 40 голосовых команд, которые являются наиболее часто используемыми командами при работе с графическим пользовательским интерфейсом русскоязычного варианта операционной системы семейства Microsoft Windows. Теоретически возможно работать с компьютером, используя лишь левую и правую кнопку мыши (команды «левая» и «правая»), однако введение дополнительных голосовых команд позволяет серьезно ускорить и упростить процесс человеко-машинного взаимодействия и производительность работы.

Подаваемые голосовые команды захватываются встроенным в видеокамеру дистанционным микрофоном, передаются в цифровом виде в компьютер и интерпретируются автоматической системой распознавания русской речи. Помимо этой системы для управления курсором мыши используется технология компьютерного зрения, отслеживающая перемещения головы пользователя. Координаты курсора мыши «привязываются» к координатам кончика носа и иных лицевых органов, автоматически определяемых на изображении, таким образом любые движения головы вызывают смещение курсора на экране в соответствующем направлении. Объединение такого интерфейса с речевым интерфейсом позволяет пользователям не только бесконтактно работать с графическим интерфейсом компьютера, подавая отдельные команды голосом, но и набирать текст в любом существующем редакторе или форме, используя виртуальную клавиатуру или произнося текст по буквам. Альтернативно для ввода текста может использоваться интерфейс Dasher [7], который является сторонней ассистивной программой для указательного ввода букв на разных языках.

2.1 Автоматическая обработка аудиовизуальных сигналов

Интерфейс способен обрабатывать одноканальный аудиосигнал от микрофона и распознавать голосовые команды на русском языке, разработаны также аналогичные версии для английского и французского языка. Для распознавания русской речи применяется оригинальная система автоматического распознавания речи, получившая название SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) [8]. В системе для параметризации звука используется разновидность спектральной обработки сигнала — мел-частотные кепстральные коэффициенты с их первой и второй производными. Акустическое моделирование звуков речи в системе производится с применением непрерывных скрытых марковских моделей (СММ) первого порядка [9] и смесей нормальных (гауссовских) распределений плотностей вероятностей векторов наблюдений в состояниях СММ. Для лучшего учета вариативности разговорной речи каждое слово преобразуется в последовательность произносимых фонем (звуков речи) и строится вероятностная модель для каждой фонемы. С помощью алгоритма Витерби вычисляется вероятность принадлежности последовательности векторов наблюдений СММ некоторого слова [9].

Для задачи голосового управления (работа с персональным компьютером относится к этой катего-

рии приложений), где применяется малый словарь распознавания, лексикон системы представляет собой линейный список всех команд с их фонематическими транскрипциями и может достаточно просто дополняться. Все голосовые команды ICanDo можно условно разделить на четыре класса по их функциональному назначению:

(1) команды, заменяющие управление кнопками и регуляторами манипулятора-мыши (например, «левая», «правая», «двойной клик», «прокрутка вниз» и т. д.);

(2) команды, заменяющие нажатие клавиш клавиатуры (например, «ввод», «удалить», «регистр», цифры, буквы и т. д.);

(3) команды управления графическим пользовательским интерфейсом (например, «открыть», «сохранить», «печать», «пуск» и т. д.);

(4) специальные команды («калибровка»).

Нужно отметить, что лишь команды, заменяющие работу мыши, являются фактически многомодальными, так как они используют информацию о положении курсора мыши в текущий момент времени. Все остальные являются исключительно речевыми одномодальными командами, и при их выполнении положение курсора не учитывается.

В многомодальном интерфейсе для управления курсором мыши используется подсистема компьютерного зрения, отслеживающая указательные движения головы пользователя. Применяется программный модуль для отслеживания движений головы пользователя, реализованный на основе базового алгоритма Лукаса–Канаде (Lukas–Kanade) [10] и его более поздней пирамидальной модификации [11] для анализа оптического потока, т. е. изображение видимого движения объектов, поверхностей или краев сцены, получаемое в результате перемещения наблюдателя относительно сцены или, наоборот, сцены относительно наблюдателя. В системе производится автоматическое отслеживание пяти естественных точек на лице: центр верхней губы, кончик носа, точка между глаз, зрачок правого глаза и зрачок левого глаза. Первоначальный поиск головы человека на статических изображениях (последовательных видеокадрах с разрешением 640×480 пикселей и частотой до 25 кадров в секунду, получаемых от видеокамеры) производится методом AdaBoost с применением алгоритма Виолы–Джонса (Viola–Jones) [12]. Изображение сканируется рамкой-окном заданного размера и строится пирамида копий объектов. Построенная пирамида анализируется заранее обученными каскадами Хаара, и на изображении находятся графические области, отвечающие заданной визуальной модели [13]. Реализованный метод детекции головы находит прямоугольные графические области на изображении, с высокой степенью вероятности содержащие изображение лица человека. Введено ограничение: размер такой области должен быть не менее 220×250 пикселей, чтобы захватывать только одно лицо в кадре, достаточно близко расположенное по отношению к видеокамере, а кроме того, это ускоряет процесс обработки видеопотока.

В отличие от имеющей аналогичное предназначение канадской системы Nouse [14], в которой отслеживается только положение кончика носа для управления движением курсора мыши, в ICanDo для более робастного слежения за перемещением головы оператора используется набор из 5 естественных лицевых объектов.

2.2 Синхронизация сигналов и объединение информации

В интерфейсе для объединения информации и выполнения многомодальной команды необходимо учитывать координаты указателя мыши, актуальные для момента времени непосредственно перед произнесением голосовой команды пользователем, т. е. должна сохраняться определенная история координат положения курсора. Если же использовать координаты указателя, актуальные на момент окончания произнесения голосовой команды, то многомодальная команда может оказаться неверной, так как курсор может сместиться от запланированного положения из-за непроизвольных перемещений головы (а они всегда существуют при говорении). В этом аспекте состоит принципиальное отличие указателя, управляемого движениями головы, от управляемого аппаратными манипуляторами наподобие мыши, трекбола, сенсорного экрана и т. д.

Звуковой сигнал, непрерывно захватываемый дистанционным стационарным микрофоном и передаваемый в компьютер посредством звуковой платы, обрабатывается модулем автоматического распознавания речи. Процесс распознавания речи запускается программным модулем детекции границ речи, который обнаруживает наличие в звуковом сигнале речевого фрагмента, отличного от тишины или постоянного фонового шума. Процесс распознавания заканчивается после получения наилучшей гипотезы распознавания голосовой команды из автоматической системы. Синхронизация модальностей производится следующим образом: текущее положение курсора сохраняется в буфере системы в первый момент определения наличия речи оператора (срабатывания алгоритма поиска границ речи по значению энергии сегментов

сигнала). По окончании процесса распознавания команды модуль распознавания речи дает сигнал на объединение информации и выполнение многомодальной команды. Таким образом, именно модуль распознавания речи осуществляет синхронизацию модальностей в бимодальном интерфейсе.

Для объединения информации, поступающей от двух модальностей, используется фреймовый метод позднего объединения, когда поля определенной структуры (фрейма) заполняются данными по мере их поступления, а по окончании процесса распознавания выполняется многомодальная команда. Поля семантического фрейма следующие: текст голосовой команды, абсцисса точки положения указателя мыши, ордината точки положения указателя, тип речевой команды (многомодальная или одномодальная). Если распознанная команда является многомодальной, она объединяется в единую команду с сохраненными координатами курсора и автоматически отсылается сообщение виртуальному устройству мыши о выполнении нужного действия. Если же голосовая команда одномодальна, то посылается соответствующее сообщение виртуальному устройству клавиатуры с кодом клавиши или сочетанием кодов. Движения головы сами по себе не могут подавать команд управления графическим пользовательским интерфейсом, однако они могут использоваться, например, для создания изображений в графических редакторах.

3 Когнитивные исследования пользовательского интерфейса

При помощи многомодального интерфейса ICanDo был проведен ряд экспериментов, которые были ориентированы на изучение организации бесконтактного взаимодействия человека с машиной и использовали элементы когнитивных исследований.

3.1 Методика исследований

Экспериментально была проведена оценка скорости и производительности работы пользователей с бесконтактным интерфейсом при указании на объекты графического пользовательского интерфейса. Для оценки скорости ввода информации была использована методология международного стандарта ISO 9241-9:2000 “Requirements for non-keyboard input devices” (Требования к неклавитаурным устройствам ввода информации) [15], которая базируется на экспериментах и законах, разработанных в середине XX в. американским психологом-когнитивистом Полом Фиттсом (Paul Morris Fitts) и впоследствии развитых другими учеными [16]. Применяемая в данном исследовании методика оценки интерфейса состоит в следующем. Тестеры, используя предоставленное им устройство указательного ввода, должны насколько возможно быстро отмечать на экране набор целей-объектов (последовательно кликнуть на них, т. е. дать голосовую команду «левая» для нажатия левой кнопки мыши), появляющихся по круговой схеме на мониторе. При этом порядок целей задается программой автоматически таким образом, чтобы пользователь последовательно выделял наиболее удаленно расположенные друг от друга объекты, совершая движения указателем в различных направлениях [17]. Когда нажатием на кнопку происходит подтверждение выделения текущего объекта-цели на экране, отображается следующая цель. При этом автоматически вычисляется индекс сложности задачи ID (*index of difficulty*), измеряемый в битах согласно формуле [18]:

$$\mathrm{ID} = \log_2\left(\frac{D}{W} + 1\right), \qquad (1)$$

где D — расстояние между центрами целей; W — диаметр цели.

Однако координаты точки, где происходит щелчок кнопкой мыши, зависят как от фактического (*effective*) расстояния между точками кликов, так и от фактического диаметра самих целей (т. е. чем меньше цель, тем сложнее попасть по ее центру). Поэтому фактический индекс сложности выражается следующей формулой [18]:

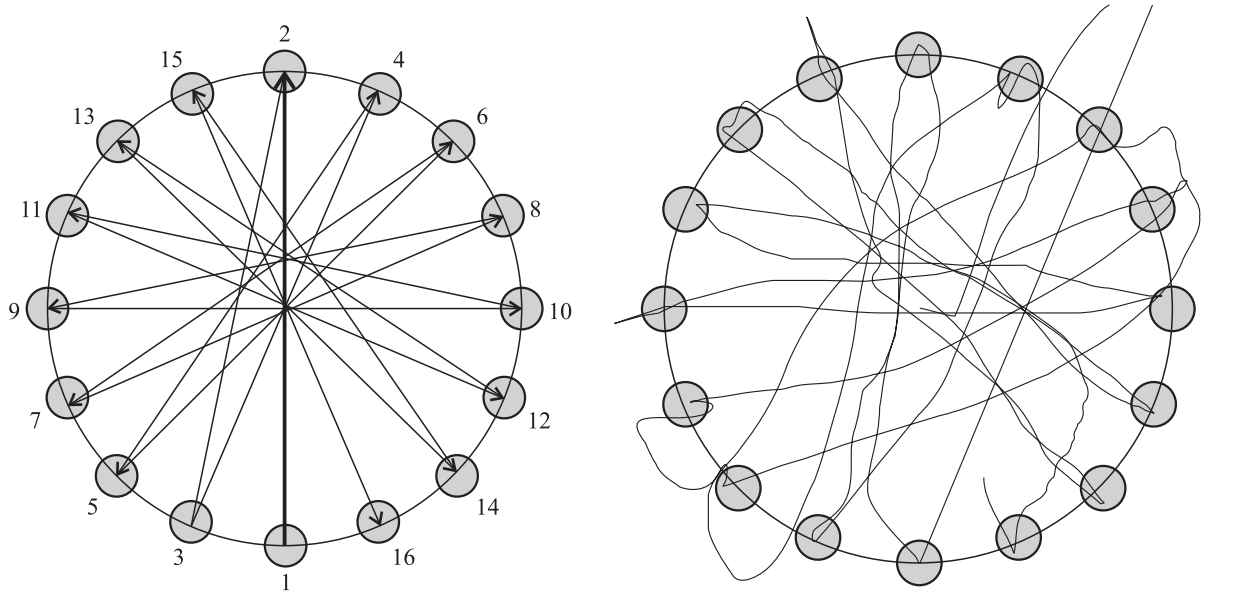
$$\mathrm{ID}_e = \log_2\left(\frac{D_e}{W_e} + 1\right). \qquad (2)$$

Здесь D_e — фактическое расстояние между точками кликов двух последних целей; W_e — фактический диаметр (или ширина) цели, определяемый в [18] как

$$W_e = 4{,}133\sigma, \qquad (3)$$

где σ — среднеквадратическое отклонение координат точки выделения (клика), проецируемой на ось, которая соединяет центры начальной и конечной целей. Получаемые значения ID_e отличаются от значений ID, более точно учитывая качество выполнения тестового задания пользователем.

Для проведения эксперимента было разработано соответствующее программное обеспечение, которое позволяет произвольно задавать значения D и W, а также фиксировать результаты прохождения теста. Программа для ЭВМ предлагает пользователю последовательно кликнуть на 16 целей, которые по очереди появляются на экране согласно



(*а*) (*б*)

Рис. 2 Схема и порядок расположения целей на экране для проведения когнитивных экспериментов с интерфейсом по методу Фиттса (*а*) и реальный пример траектории движения курсора мыши на экране при бесконтактном выполнении задания (*б*)

рис. 2, *а*. На рис. 2, *б* показан реальный пример получившейся траектории движения курсора мыши на экране при бесконтактном выполнении задания посредством ICanDo. Здесь можно видеть, что данная задача для пользователя не была простой, но ошибок выделения (непопаданий по целям) он не допустил.

3.2 Анализ результатов экспериментов

Для выполнения тестового задания были привлечены четыре пользователя-новичка, не имевших ранее опыта работы с многомодальным интерфейсом, и два пользователя-эксперта, принимавших участие в ее разработке и отладке. Каждым пользователем были проведены серии по 10 тестов с последовательным изменением диаметра цели W в пределах от 32 до 128 пикселей и среднего расстояния D между целями в пределах 96–650 пикселей (использовалось разрешение экрана 1280×1024), т. е. показатель ID варьировался от 1,32 до 4,4 бит. Каждый тест занимал в среднем 30–60 с.

Рисунок 3 показывает полученный в результате экспериментов и усредненный по всем пользователям график зависимости отношения значений ID_e (фактический индекс сложности) и ID (теоретически рассчитанный индекс сложности) при разных значениях D и W. Характерно, что данный график лежит выше пунктирной линии-нормали (ожидаемый теоретически индекс сложности выполнения задачи), а это означает, что выполнение данной задачи оказалось несколько сложнее, чем планировалось. В противном случае, если бы график зависимости лежал ниже нормали, то можно было бы говорить о том, что предлагаемая тестерам задача оказалась легче расчетной сложности.

Согласно экспериментам по методике Фиттса, время движения MT (*movement time*) между двумя целями линейно зависит от индекса сложности ID [19]. Полученное в ходе экспериментов среднее значение MT для всех тестеров равнялось 2550 мс, т. е. около 2,5 с между речевыми «нажатиями» цели. Рисунок 4 показывает два аппроксимирующих графика зависимости времени движения MT от фактической сложности задачи ID_e отдельно для пользователей-новичков, не работавших ранее с интерфейсом, и для обученных пользователей-экспертов. Хорошо заметно, что эффект обучения положительно сказывается на увеличении скорости бесконтактной работы с компьютером. Также разброс значений MT для новичков оказался значительнее, они выполняли тесты менее стабильно. На основании результатов экспериментов можно сказать, что новички начинают уверенно работать с компьютером бесконтактно при помощи многомодального интерфейса уже через 10–15 мин тренировки (исключая этап настройки системы распознавания речи на голосовые характеристики

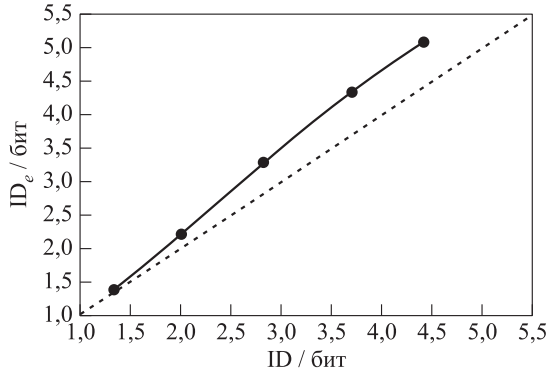


Рис. 3 График зависимости значений фактической сложности ID_e и теоретической сложности ID выполнения задачи и его отклонение от нормали

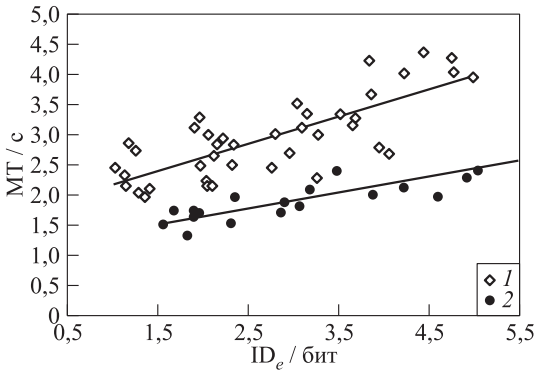


Рис. 4 Графики зависимостей времени движения MT от фактического индекса сложности ID_e задачи отдельно для новичков (*1*) и экспертов (*2*)

пользователя), что, конечно же, несколько больше, чем при первоначальном овладении мышкой и клавиатурой. Однако через день работы с системой пользователь уже может считаться экспертом в бесконтактном человеко-машинном взаимодействии.

В применяемой методике экспериментов Фиттса основным показателем оценки интерфейса является общая производительность работы пользователя с системой TP (*throughput*) [20], определяющая компромисс между временем движения (скоростью выполнения задания) и точностью выделения цели и измеряемая в битах в секунду согласно следующей формуле:

$$\mathrm{TP} = \frac{\mathrm{ID}_e}{\mathrm{MT}}. \qquad (4)$$

Полученное в ходе экспериментов среднее значение TP для всех тестеров составило 1,2 бит/с, максимальное значение TP для одного тестера — 2,0 бит/с.

Также в ходе когнитивных исследований была проведена сравнительная оценка контактных устройств для ввода/указания, таких как сенсорный экран 17″, джойстик, трекбол, сенсорная панель (*touchpad*) 3″ и стандартный манипулятор-мышь. Двумя пользователями были проведены серии по 10 тестов для каждого устройства с последовательным изменением диаметра цели W в пределах от 32 до 128 пикселей и среднего расстояния D между целями в пределах от 96 до 650 пикселей. Таблица 1 приводит результаты экспериментов и сравнения всех вышеуказанных устройств по трем основным количественным критериям:

(1) среднее время движения MT между двумя целями;

(2) процент ошибок выделения целей (непопадание курсором в цель);

(3) общая производительность указательного интерфейса TP.

Таблица 1 показывает, что наилучшие результаты по производительности интерфейсов были показаны сенсорным монитором, так как рука тестера свободно перемещается по воздуху. Управление курсором посредством многомодального интерфейса, отслеживающего движения головы, уступает по производительности практически всем аппаратным контактным средствам ввода информации, кроме джойстика (который весьма непривычен для управления курсором), однако имеет то преимущество, что является бесконтактным способом управления курсором и может применяться категориями потенциальных пользователей, для которых стандартные средства ввода информации недоступны.

Таблица 1 Сравнительная оценка эффективности интерфейсов для указательного ввода информации с использованием методики Фиттса

Устройство ввода	MT, с	Ошибка выделения, %	TP, бит/с
Джойстик	2,01	7,00	1,54
Трекбол	1,03	3,83	3,51
Сенсорная панель 3″	0,85	4,50	3,72
Манипулятор-мышь	0,49	3,17	6,65
Сенсорный экран 17″	0,50	6,17	7,85
Интерфейс ICanDo	1,98	7,33	1,59

Тестирование интерфейса в реальной задаче бесконтактной работы с компьютером было также проведено тремя добровольными пользователями. Пользователям предлагался определенный сценарий — последовательность операций, которую

Таблица 2 Сравнение бесконтактного и контактного интерфейсов человеко-машинного взаимодействия

Точность распознавания голосовых команд, %	Время выполнения тестового сценария, с	
	Интерфейс ICanDo	Мышь + клавиатура
96	82	43

пользователи должны были выполнить двумя способами (многомодальным — посредством ICanDo и стандартным — при помощи манипулятора-мыши). Тестовая задача включала в себя элементарные операции с текстовым редактором MS Word, а также поиск заданной информации в Интернете посредством MS Internet Explorer. Конкретнее: пользователю нужно было найти информацию о программе передач на интернет-портале Рамблер, скопировать интересующий фрагмент этой страницы, открыть текстовый редактор MS Word, вставить в пустой документ информацию из буфера, сохранить файл на рабочем столе и распечатать данный файл. Таблица 2 показывает количественные результаты экспериментов и сравнение двух способов человеко-машинного взаимодействия (среднее время, требуемое для выполнения всего тестового сценария и точность распознавания речи в дикторозависимом режиме).

Многомодальный бесконтактный способ ввода оказался в 1,9 раз медленнее, чем стандартный контактный способ, что было очевидно. При этом точность распознавания голосовых команд составила свыше 96% в дикторозависимом режиме работы. Однако, если учесть, что аудиосигнал, получаемый от встроенного в видеокамеру микрофона, характеризуется невысоким отношением сигнал/шум (SNR, signal-to-noise ratio), то полученный результат по точности распознавания можно считать приемлемым. Полученная скорость работы бесконтактного интерфейса вполне достаточна, так как он разрабатывается для помощи людям с физическими ограничениями, в частности для людей без рук или с парализованными руками.

Также был проведен анализ статистики бесконтактной работы пользователя-эксперта с интерфейсом ICanDo в течение одного дня в задаче навигации (серфинга) в Интернете посредством браузера MS Internet Explorer. Последующий анализ журнала статистики показал, что всего пользователь сделал более 750 голосовых команд, при этом некоторые команды были более частотными, чем другие, а некоторые команды не использовались вовсе. Диаграмма на рис. 5 показывает распределение частотности голосовых команд, примененных пользователем.

Легко было предсказать заранее, что наиболее популярной окажется команда «левая» (клик левой кнопкой мыши), которая использовалась более чем в трети случаев, включая и ввод текста при помощи специального программного обеспечения — экранной виртуальной клавиатуры. Однако необходимо сказать, что при работе с мышкой и клавиатурой это значение еще выше для подобной задачи, так как, работая бесконтактно, пользователи стараются избегать работы со сложными многоуровневыми меню стандартных офисных прикладных программ, заменяя их «горячими клавишами» для быстрого доступа к действиям. Все остальные команды рас-

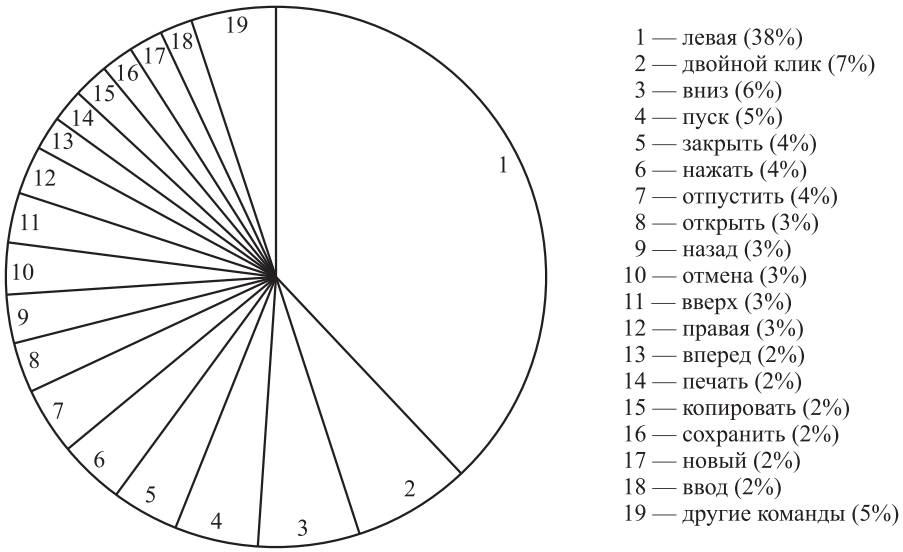


Рис. 5 Распределение относительной частоты использования голосовых команд тестером в ходе эксперимента

пределены более-менее равномерно среди оставшихся 62%. При этом 64% всех голосовых команд было подано многомодально (совместно с движениями головы для выделения графических объектов или ссылок на экране), а оставшиеся 36% команд — одномодально.

Видеодемонстрации бесконтактной работы пользователей с компьютером, в том числе и одного человека, не имеющего верхних конечностей, посредством ассистивного многомодального интерфейса ICanDo можно посмотреть на интернет-сайте лаборатории речевых и многомодальных интерфейсов СПИИРАН [21].

4 Заключение

В статье представлены результаты исследований бесконтактного человеко-машинного взаимодействия, реализуемого посредством ассистивного многомодального интерфейса ICanDo, предназначенного специально для работы человека-оператора с ЭВМ без использования рук. Описана общая архитектура ассистивного многомодального интерфейса, автоматическая обработка аудио- и видеосигналов, а также механизмы синхронизации и объединения модальностей. В данном ассистивном пользовательском интерфейсе для робастного отслеживания указательных жестов/движений головы оператора используется массив из пяти естественных точек на лице: центр верхней губы, кончик носа, точка между глаз, зрачок правого глаза и зрачок левого глаза. Применяются голосовые команды для бесконтактного управления прикладным и системным программным обеспечением компьютера. Результаты проведенных исследований с использованием методики Фиттса и элементов иных когнитивных экспериментов позволяют заключить, что данный многомодальный интерфейс обеспечивает приемлемую скорость и производительность работы пользователя с компьютером, не сильно отличающуюся от аналогичных показателей для стандартных контактных интерфейсов — устройств ввода, и может успешно применяться для бесконтактного управления как обычными операторами, так и потенциальными пользователями-инвалидами с грубыми моторными нарушениями в функционировании рук и даже вовсе без верхних конечностей.

Применение ассистивного пользовательского интерфейса позволит повысить социоэкономическую интеграцию инвалидов в информационном обществе и сделает их более независимыми от помощи со стороны других лиц. Предложенный бесконтактный интерфейс позволит пользователям самим выбирать доступные им средства взаимодействия с компьютером, компенсируя недоступные модальности альтернативными коммуникативными каналами.

Литература

1. *Карпов А. А.* ICanDo: Интеллектуальный помощник для пользователей с ограниченными физическими возможностями // Вестник компьютерных и информационных технологий, 2007. № 7. С. 32–41.
2. *Кричевец А.* Шлемомышь // Компьютерра, 2002. № 434. С. 48–51. www.computerra.ru/offline/2002/434/16588/.
3. *Bates R., Istance H. O.* Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices // 1st Cambridge Workshop on Universal Access and Assistive Technology Proceedings. — USA, 2002.
4. *Аграновский А. В., Евреинов Г. Е., Яшкин А. С.* Аппаратно-программные инструментальные средства проектирования виртуальных акустических объектов и сцен для слепых пользователей персональных компьютеров // Информационные технологии в образовании: Мат-лы IX Междунар. конф.-выставки. — М., 1999.
5. *Карпов А. А., Ронжин А. Л.* Многомодальные интерфейсы в автоматизированных системах управления // Известия высших учебных заведений. Приборостроение, 2005. Т. 48. № 7. С. 9–14.
6. *Ронжин А. Л., Карпов А. А.* Проектирование интерактивных приложений с многомодальным интерфейсом // Докл. Томского гос. ун-та систем управления и радиоэлектроники (ТУСУР), 2010. № 1. Ч. 1. С. 124–127.
7. *Ward D., Blackwell A., MacKay D.* Dasher: A data entry interface using continuous gestures and language models // ACM Symposium on User Interface Software and Technology UIST'2000 Proceedings. — New York: ACM Press, 2000. P. 129–137.
8. *Ronzhin A. L., Karpov A. A.* Russian voice interface // Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications), 2007. Т. 17. № 2. С. 321–336.
9. *Карпов А. А.* Аудиовизуальный речевой интерфейс для систем управления и оповещения // Известия Южного федерального ун-та. Технические науки, 2010. № 3(104). С. 218–222.
10. *Lucas B. D., Kanade T.* An iterative image registration technique with an application to stereo vision // 7th Joint Conference (International) on Artificial Intelligence IJCAI Proceedings. — Vancouver, Canada, 1981. P. 674–679.
11. *Bouguet J.-Y.* Pyramidal implementation of the Lucas–Kanade feature tracker description of the algorithm // Intel Corporation Microprocessor Research Labs: Report. — New York, USA, 2000.

12. *Viola P., Jones M.* Rapid object detection using a boosted cascade of simple features // IEEE Conference (International) on Computer Vision and Pattern Recognition Conference (CVPR) Proceedings. — Kauai, HI, USA, 2001.
13. *Lienhart R., Maydt J.* An extended set of Haar-like features for rapid object detection // IEEE Conference (International) on Image Processing (ICIP'2002) Proceedings. — Rochester, New York, USA, 2002. P. 900–903.
14. *Gorodnichy D., Roth G.* Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces // Image and Vision Computing, 2004. Vol. 22. No. 12. P. 931–942.
15. ISO 9241-9:2000(E) Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 9: Requirements for Non-Keyboard Input Devices. — International Standards Organization, 2000.
16. *Soukoreff R. W., MacKenzie I. S.* Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI // Intern. J. Human Computer Studies, 2004. Vol. 61. No. 6. P. 751–789.
17. *Zhang X., MacKenzie I. S.* Evaluating eye tracking with ISO 9241 Part 9 // Human–Computer Interaction Conference (International) (HCII 2007) Proceedings. — Beijing, China: Springer Verlag LNCS 4552, 2007. P. 779–788.
18. *Carbini S., Viallet J. E.* Evaluation of contactless multimodal pointing devices // 2nd IASTED Conference (International) on Human–Computer Interaction Proceedings. — Chamonix, France, 2006. P. 226–231.
19. *De Silva G. C., Lyons M. J., Kawato S., Tetsutani N.* Human factors evaluation of a vision-based facial gesture interface // Workshop on Computer Vision and Pattern Recognition for Computer Human Interaction Proceedings. — Madison, USA, 2003.
20. *Wilson A., Cutrell E.* FlowMouse: A computer vision-based pointing and gesture input device // Human–Computer Interaction INTERACT Conference Proceedings. — Rome, Italy, 2005. P. 565–578.
21. Видеодемонстрации с интернет-сайта лаборатории речевых и многомодальных интерфейсов СПИИРАН. www.spiiras.nw.ru/speech/demo/demo_new.avi, www.spiiras.nw.ru/speech/demo/ort.avi.