Речевой интерфейс в управлении текстовым редактором MS Word

Бондаренко И.Ю., Федяев О.И.
Донецкий национальный технический университет
fedyaev@r5.dgtu.donetsk.ua

Abstract

Bondarenko I., Fedyaev O. Speech interface in control of MS Word text editor. *Inclusion of the speech interface in text editor for the purpose of increase of the user's work efficiency during creating and editing of documents is considered. The advanced program model of process control of the text input and editing based on integration of graphical and voice user interface means is offered.*

Введение

Современное состояние интерфейса компьютеризированных систем не позволяет решать новые задачи человеко-машинного взаимодействия. Один из подходов к разработке эргономичных пользовательских интерфейсов основан на человеко-машинном взаимодействии с помощью устной речи как на наиболее естественном для человека способе общения [1]. По прогнозам Б. Гейтса, речевое управление, наряду с видеосистемами, распознающими двигательную активность оператора, и сенсорными датчиками, в течение ближайших десяти лет станет основой пользовательского интерфейса компьютерных систем [2].

Особо актуальной сейчас является задача построения речевого интерфейса для управления процессом ввода и редактирования текстовой информации, поскольку работа с электронным текстом наиболее широко распространена среди большинства пользователей, в том числе и неподготовленных в области информатики, а современный текстовый редактор – это сложная многофункциональная система, требующая от человека-оператора специальной подготовки.

Можно выделить 3 подхода к организации взаимодействия человека и компьютера при обработке текстовых документов:

1. *Набор текста с клавиатуры и его редактирование с помощью графического интерфейса.* При этом вся нагрузка по человеко-машинному взаимодействию ложится на тактильно-зрительный канал. Человек делает опечатки, отвлекаясь от основной цели – набора текста – на поиск пунктов графического меню, вспоминает «горячие клавиши» и т.п., и в результате устаёт. Речевой канал взаимодействия при этом бездействует.

2. *Автоматическое преобразование речи пользователя, диктующего документ, в текст.* Человек зачитывает компьютеру весь вводимый текст, имеющий зачастую весьма большой объём. Скорость речи и скорость набора текста на языках с алфавитным письмом примерно одинаковы. Кроме того, для автоматической диктовки необходимо не просто распознавать голосовые сигналы, а понимать слитную речь так, как это делает человек, – в контексте окружающей среды, опыта и внутренней картины мира собеседника. Создать такую систему – это значит создать полноценную систему искусственного интеллекта, аналогичную системе «естественного» интеллекта человека. Даже принципиальная возможность решения подобной задачи многими учёными ставится под сомнение. Современные системы речевой диктовки обеспечивают очень низкое качество ввода текста и непригодны для практического применения (исключение составляют лишь системы речевой диктовки на языках с иероглифическим письмом, где неудобства клавиатурного ввода иероглифов заведомо превосходят неудобства, связанные с исправлением ошибок систем речевой диктовки).

3. *Сочетание клавиатурного ввода текста и его редактирования посредством голосовых команд.* На наш взгляд, в современных условиях именно этот вариант является оптимальным с точки зрения эргономики. В данном случае пользователь при вводе текста не переключает фокус внимания на графические меню и не пытается вспомнить сочетания «горячих клавиш», которые трудно запоминаются большинством людей в силу своей неочевидности. Вместо этого для редактирования вводимого текста используются естественные и легко запоминаемые речевые команды. При этом нагрузка равномерно распределяется между тактильно-зрительным и речевым каналами взаимодействия, что снижает утомляемость пользователя и повышает производительность его труда.

Таким образом, целью данной работы является разработка речевого интерфейса текстового редактора и его включение в общую структуру управления процессом ввода и редактирования текстовой информации. Для достижения данной цели необходимо проанализировать особенности существующих средств речевого управления текстовыми редакторами, разработать структуру и программную модель речевого интерфейса.

Обзор существующих средств речевого управления текстовыми редакторами

Компания Dragon Systems была первой компанией, предложившей в 1997г. программу речевого управления текстовым редактором общего пользования. Сейчас эта компания предложила улучшенный вариант программы – Dragon NaturallySpeaking 9.0 [3]. Этот програмный пакет для автоматической диктовки включает большой набор команд форматирования и редактирования, хотя предполагает использование только собственного текстового процессора.

Главным конкурентом Dragon NaturallySpeaking является программа от IBM ViaVoice [4]. Она имеет ряд преимуществ, среди которых стоит выделить хорошо проработанную систему команд и возможность интеграции с Microsoft Word. Однако тесты показали, что она не столь точна, как NaturallySpeaking и имеет мало голосовых команд редактирования и навигации [5].

Компанией Philips разработана система речевой диктовки SpeechMagic и её узкоспециализированная версия, предназначенная для ввода и редактирования медицинских текстов. По опубликованным данным, система имеет активный словарь из 64 000 слов и базовый словарь из 270 000 слов с возможностью расширения его пользователем [6].

Корпорация Microsoft также активно занимается внедрением речевого интерфейса в свои программные продукты. С помощью компонент MS Speech API программист может организовывать речевой интерфейс в любой прикладной программе. Заявленная точность распознавания составляет порядка 95% [7].

Несмотря на достаточно большое число подобных систем речевого управления и диктовки, пользователи до сих пор не сделали однозначный выбор какого-либо конкретного способа и системы речевого управления, что, на наш взгляд, связано с невысоким качеством распознавания речевых команд. Все вышеописанные системы речевого управления и диктовки, по заявлениям их разработчиков, обладают точностью распознавания свыше 90%. Однако методики проведения экспериментов, в результате которых было продемонстрировано столь высокое качество распознавания устной речи, не приведены ни одной из фирм-разработчиков. Независимое тестирование, напротив, показывает невысокую точность и надёжность работы существующих систем распознавания речи. В частности, IBM ViaVoice и Dragon NaturallySpeaking показали невысокие результаты в тестах, описанных в статьях раздела «Тестовая лаборатория» WEB-портала речевых технологий «Голосовые технологии», не достигнув точности даже 90% [5]. Следовательно, необходимость разработки удобного и надёжного речевого интерфейса текстового редактора по-прежнему является актуальной.

Структура речевого интерфейса

Ключевой особенностью управления любой технической системы является то, что она в ответ на неограниченное число ситуаций внешнего мира предусматривает выполнение строго ограниченного числа действий. Это позволяет без снижения эргономичности управления представить голосовые команды не как контекстно-свободные фразы слитной речи, а как конечный набор изолированных речевых сигналов, поступающих от пользователя. Таким образом, при разработке голосового канала управления задача распознавания голосовых команд сводится к задаче распознавания изолированных речевых сигналов из ограниченного множества.

В условиях, когда словарь системы распознавания изолированных речевых сигналов не является сверхбольшим, наибольшую точность показывают методы целостного распознавания, основанные на распознавании речевого сигнала как целого слова, без разбиения его на более мелкие речевые единицы (фонемы, аллофоны и т.п.) [8]. Одним из таких методов, показавшим высокие результаты при распознавании слов японской, английской и немецкой речи, является нечёткое сопоставление образов [9]. Речевые сигналы в этом методе путём спектрального анализа преобразуются в двумерные спектрально-временны*е* образы (СВО). Затем для каждого временно*г*о среза СВО определяются номера частот, на которых произошли амплитудные всплески, называемые ещё локальными выбросами [9]. Эти номера кодируются единицами, а остальные – нулями, что позволяет отразить изменение структуры локальных выбросов в спектре речевого сигнала с течением времени. Получённые двоичные спектрально-временные образы (ДСВО) подаются на вход системы распознавания. Эталоны речевого словаря представляются в виде нечётких отношений между номерами частот, на которых обнаружены локальные выбросы, и номерами временны*х* интервалов спектрального анализа. Данный метод был адаптирован также и для распознавания слов русской речи [10].

При сопоставлении входного двумерного образа и нечёткого образа-эталона необходимо выполнить процедуру временно*й* нормализации, т.е. привести эти образы к одной длине по оси времени, поскольку различные реализации речевых образов, даже относящихся к одному и тому же классу, могут значительно отличаться друг от друга по длительности. Основной проблемой нормализации, сильно влияющей на качество распознавания, является временна*я*

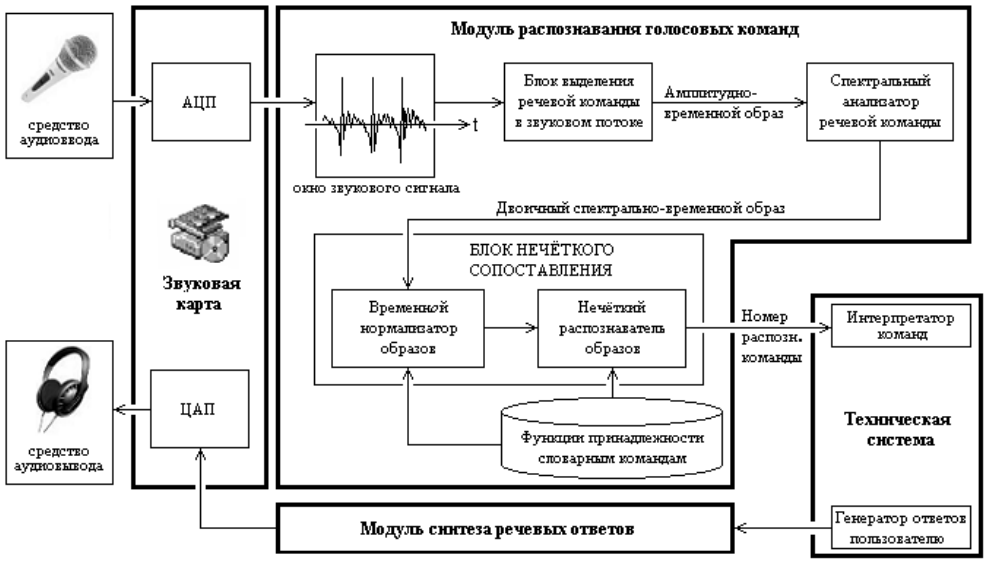

Рисунок 1 – Структура канала голосового управления компьютерными системами

нестабильность речевого сигнала, т.е. неравномерность протекания речевого сигнала во времени, вызванная перманентной нестабильностью темпа речи пользователя, влиянием интонации, акцента и т.п. С этой проблемой сталкиваются все методы распознавания речевых сигналов.

Были предложены различные пути решения проблемы временной нестабильности речевого сигнала, основанные на линейном [9, 12], нелинейном [12] и оптимальном [13, 14] временном выравнивании речевых образов. Результаты экспериментальных исследований, проведённых на речевой многодикторной базе данных, включавшей в себя команды управления текстовым редактором, показали преимущество последнего подхода [13, 14].

Спроектирована структурная схема канала голосового управления компьютерными системами с использованием метода нечёткого сопоставления образов для распознавания управляющих команд оператора (рис.2), а также разработана специализированная программная система русскоязычного голосового управления процессом ввода и редактирования текстовой информации, интегрированная в общий контур управления текстовым редактором Microsoft Word. В качестве алгоритма распознавания речевых команд в данной системе был использован метод нечёткого сопоставления образов с оптимальным временным выравниванием.

Программная модель речевого интерфейса

Речевой интерфейс текстового редактора MS Word реализован в виде специализированной программной системы, работающей в фоновом режиме и автоматически запускающейся при включении компьютера. Первая версия такой системы была описана в работе [7]. В данной работе предложена новая программная модель речевого интерфейса, которая имеет следующие отличительные характеристики:

1. Выделяется три типа речевых команд пользователя: ключевые слова, управляющие команды и команды-параметры.
   1.1. Ключевые слова используются для определения ситуации, когда пользователь обращается к речевому интерфейсу (по аналогии с тем, как человек обращается по кличке к своему домашнему животному, чтобы привлечь его внимание для передачи управляющей команды).
   1.2. Управляющие команды предназначены для непосредственной передачи управляющего воздействия (например, произнесение пользователем команды «Новая строка» означает его желание вставить пустую строку в текущую позицию документа).
   1.3. Команды-параметры конкретизируют условия, при которых должно быть передано управляющее воздействие (например, управляющее воздействие, вызываемое командой «Моя таблица», не может быть передано текстовому редактору без определения следующего условия: сколько строк и столбцов должно быть во вновь создаваемой таблице).
2. Пользователь имеет более широкие возможности по формированию набора управляющих команд, которые он имеет возможность описывать на языке Visual Basic for Applications (рис. 2).
3. Пользователь имеет возможность настраивать систему на восприятие особенностей своего голоса с помощью специальной процедуры обучения (рис. 3).

4. В журнале событий автоматически регистрируются все действия пользователя по речевому управлению текстовым редактором, а также ответные действия управляемой системы (рис. 4).

При разработке программной модели речевого интерфейса возникает задача рациональной организации вычислительного процесса. В соответствии со структурой канала голосового управления компьютерными системами (рис. 1), оцифрованный звуковой сигнал поступает с АЦП на вход системы речевого управления порциями одинаковой длительности T. После поступления очередной порции звуковых данных осуществляется их обработка модулем распознавания речевых команд. Обозначим время, затрачиваемое модулем распознавания речевых команд на выполнение необходимых вычислительных операций, как t_M . Модуль распознавания может работать в двух режимах:

- режиме поиска команды;
- режиме распознавания команды.

В первом режиме работает только блок выделения границ речевой команды, проверяющий, содержит ли вновь поступившая порция звуковых данных признаки начала речевой команды. Время работы блока составляет τ_1, и в таком случае $t_M = \tau_1$.

Во втором режиме после окончания работы блока выделения границ речевой команды, проверяющего в данном случае признаки окончания речевой команды, начинают выполняться процессы спектрального анализа и распознавания выделенной речевой команды, имеющие суммарное время выполнения τ_2 . В этом режиме $t_M = \tau_1 + \tau_2$.

Если команда идентифицирована как известная и выполнимая, то запускается процесс исполнения данной команды текстовым редактором Microsoft Word. Обозначим время выполнения этого процесса как t_S .

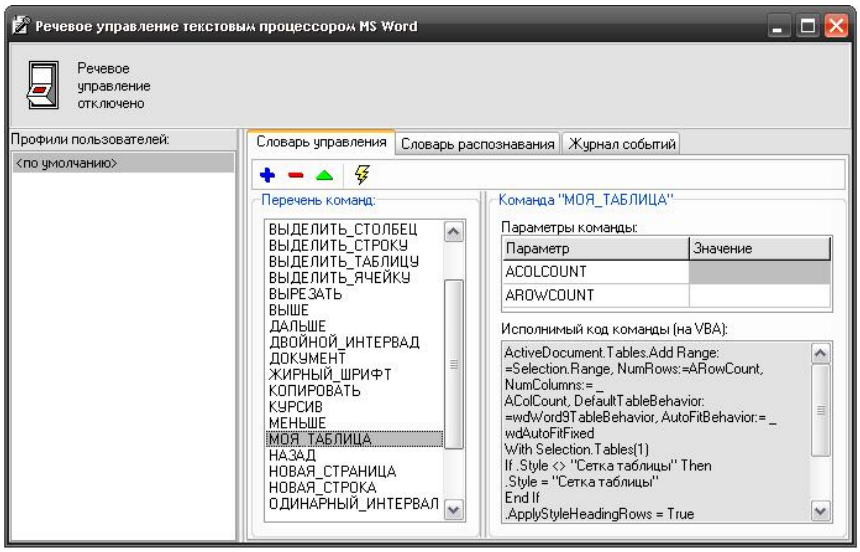

Рисунок 2 – Главное окно настройки речевого интерфейса (вкладка «Словарь управления»)

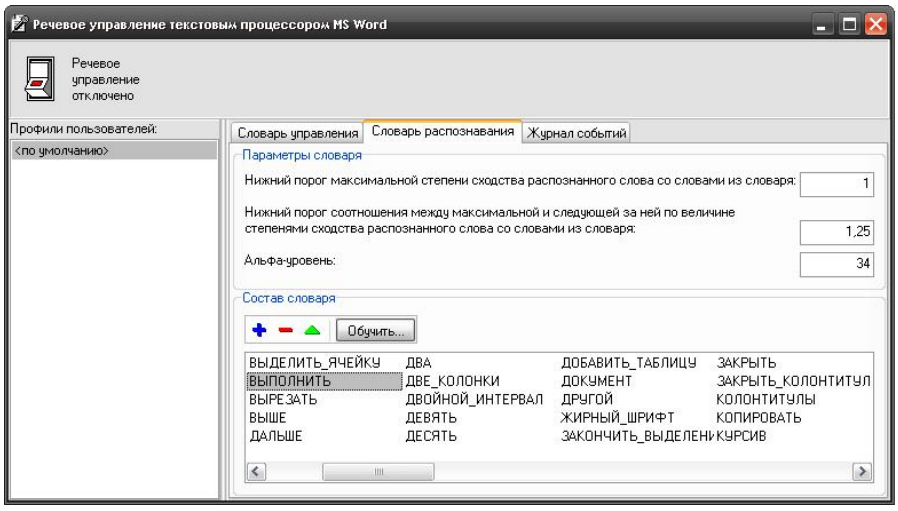

Рисунок 3 – Главное окно настройки речевого интерфейса (вкладка «Словарь распознавания»)

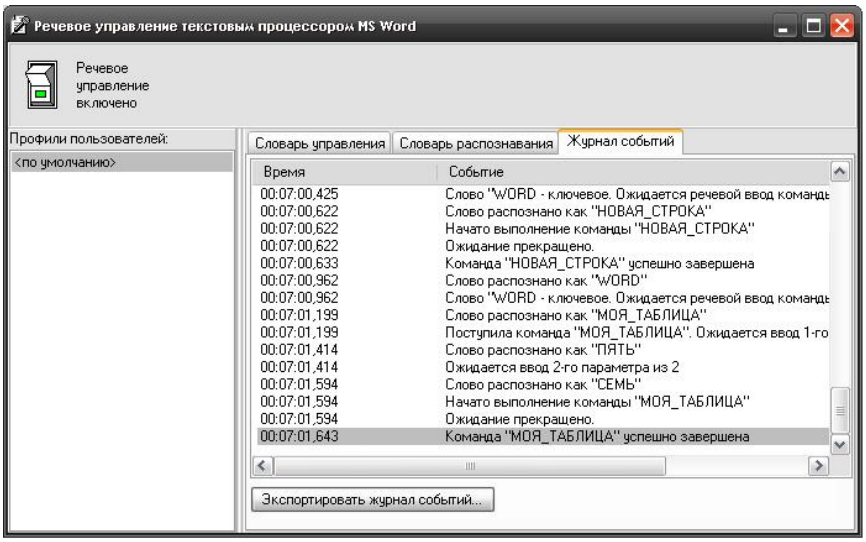

Рисунок 4 – Главное окно настройки речевого интерфейса (вкладка «Журнал событий»)

Для предотвращения потерь очередных порций звуковых данных, поступающих с АЦП на вход системы речевого управления, необходимо учесть следующее ограничение на время выполнения рассмотренного вычислительного процесса:

$$t_M + t_S \le T. \qquad (1)$$

Как было показано ранее, время распознавания t_M зависит от скорости выделения границ речевой команды (τ_1), объёма словаря распознавания и длительности выделенной команды (τ_2). Из анализа следует, что для словарей распознавания среднего объёма, как правило, $t_M < T$. В то же время выполнение команды текстовым редактором t_S может длиться очень долго, поскольку на него оказывают влияние алгоритмическая сложность макроса, соответствующего распознанной речевой команде, и скорость выполнения этого макроса интерпретатором VBA, встроенным в текстовый редактор Microsoft Word. Поэтому условие (1) может нарушаться, и необходимо таким образом спроектировать вычислительный процесс, чтобы на работу модуля распознавания речевых команд не оказывало влияние время работы управляемой системы. Решения данной задачи можно добиться, если рассматривать распознавание речевой команды и её выполнение как два независимых вычислительных процесса, выполняющихся параллельно (например, предыдущая речевая команда ещё не выполнилась, а новая речевая команда, поступившая на вход, уже начинает распознаваться). Для разрабатываемой программной модели речевого интерфейса, работающей в операционной системе Microsoft Windows, данная задача была решена путём использования принципа многопоточного программирования, где в одном потоке функционирует модуль распознавания речевых команд, а в остальных потоках осуществляется выполнение ранее распознанных команд управления текстовым редактором. Для обеспечения асинхронной работы потоков использован буфер управляющих команд, совместный доступ к которому реализован на основе стандартных механизмов ОС Windows – мьютексов и критических секций.

Выводы

В данной работе исследованы вопросы, связанные с разработкой речевого пользовательского интерфейса текстового редактора и включением этого интерфейса в общую систему управления процессом ввода и редактирования текстовой информации. В рамках проведённых исследований получены следующие результаты:

1) проанализированы недостатки современных средств управления процессом ввода и редактирования текстовой информации;

2) предложена структура речевого интерфейса текстового редактора, позволяющая распределить процесс управления вводом и редактированием текстовой информации между двумя каналами информационного обмена: тактильно-зрительным и речевым;

3) разработана программная модель речевого интерфейса в управлении текстовым редактором Microsoft Word как наиболее распространённом среди пользователей ОС Windows.

Дальнейшие исследования будут проводиться в следующих направлениях:

- исследование и развитие обобщенной архитектуры речевого интерфейса в управлении технической системой;
- разработка модели для оценки эргономичности управления технической

системой при использовании речевого пользовательского интерфейса;

- создание методов распознавания речи, повышающих надёжность функционирования речевого интерфейса в дикторонезависимом режиме.

Литература

1. Плотников В.Н. и др. Речевой диалог в системах управления. – М.: Машиностроение. – 1988. – 224 с.
2. G. Gross. Gates: Next decade will bring huge software advances. InfoWorld. March 13, 2008. http://www.infoworld.com/article/08/03/13/Gates-says-next-decade-will-bring-huge-software-advances_1.html
3. Nuance – Dragon NaturallySpeaking 9. – 17.5.2006. – http://www.nuance.com /naturallyspeaking.
4. IBM Software – IBM ViaVoice – Product Overview. – 6.04.2006. – http://www-306.ibm.com /software/voice/viavoice/
5. Информационный портал речевых технологий «Голосовые технологии». Тестовая лаборатория. – 6.04.2006. – http://art.bdk.com.ru/govor/1listr62.htm.
6. Royal Philips Electronics – Philips Speech Recognition Systems. – 6.04.2006. – http://www.speechrecognition.philips.com.
7. Буторин Д.Н. MS Agent и Speech API в Delphi. – С.-Пб.: BHV, 2005. – 448 с.
8. Жожикашвили В.А. и др. Применение распознавания речи в автоматизированных системах массового обслуживания // Автоматизация и современные технологии. – 2003. – № 11. – С.23-29.
9. Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др. Распознавание речи // Прикладные нечёткие системы. Под ред. Т.Тэрано, К. Асаи, М. Сугено. – М.: «Мир», – 1993. – С.157-170.
10. Федяев О.И., Бондаренко И.Ю. Интеграция визуального и речевого способов управления процессом ввода и редактирования текстовой информации // Сб. тр. конф. «Научная сессия МИФИ-2006». Т.3. Интеллектуальные системы и технологии. – М.: МИФИ. – 2006. – С. 194 –195.
11. Федяев О.И., Бондаренко И.Ю. Речевое управление текстовым редактором MS Word // Сб. тр. 3-й научно-методич. конф. «Проблемы и пути усовершенствования научно-методической и учебно-воспитательной работы в ДонНТУ». - Донецк: ДонНТУ. – 2007.
12. Бондаренко И.Ю, Федяев О.И. Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов // Сб. тр. VI междунар. науч. конф. «Интеллектуальный анализ информации ИАИ-2006». – 2006. – С.20 – 27.
13. Федяев О.И., Бондаренко И.Ю. Нечёткое сопоставление образов с оптимальным временным выравниванием для однодикторного и многодикторного распознавания изолированных слов // Сб. науч. трудов Донецкого нац. техн. ун-та. Серия «Информатика, кибернетика и вычислит. техника». – 2007. – Выпуск 8 (120). – С.273–281.
14. Бондаренко И.Ю., Федяев О.И. Голосовое командное управление и проблема временной нестабильности речевого сигнала // Сб. тр. VIII междунар. науч. конф. «Интеллектуальный анализ информации ИАИ-2008». – 2008. – С. 110 – 116.