УДК 681.3:519.816

ИНТЕГРАЦИЯ ВИЗУАЛЬНОГО И РЕЧЕВОГО СПОСОБОВ УПРАВЛЕНИЯ ТЕКСТОВЫМ РЕДАКТОРОМ

О.И.Федяев, С.А.Гладунов, И.Ю.Бондаренко
Донецкий национальный технический университет
fedyaev@r5.dgtu.donetsk.ua, bond005@yandex.ru

Розглядається включення мовленнєвого інтерфейса у текстовий процессор Microsoft Word з метою підвищення ефективності роботи користувача під час вводу та редагування документів. Пропонується структура мовленнєвого каналу інтерфейса на основі взаємодії сегментной та цілісної систем розпізнавання, перша з яких реалізована у нейромережевому, а друга – у нечіткому базисі.

Рассматривается проблема создания средств речевой коммуникации между человеком и компьютерными системами. Добавление речевого канала в контур управления сложными человеко-машинными системами позволит значительно повысить эффективность их работы. Такой сложной с точки зрения интерактивного взаимодействия системой является современный текстовый редактор, предоставляющий пользователю большое количество функциональных возможностей и связанных с ними команд по вводу и редактированию различного типа информации. Рациональное сочетание речевого и стандартного визуального способов управления процессом ввода и редактирования текстовой информации позволит снизить нагрузку на тактильно-зрительные каналы человека и тем самым повысить эффективность его работы. О востребованности речевого интерфейса свидетельствует и возросшее число коммерческих разработок систем, использующих речевой интерфейс. Так, NaturallySpeaking фирмы Dragon System позволяет редактировать и форматировать текст с помощью собственного текстового процессора без использования клавиатуры и мыши. Компания IBM разработала аналогичную программу, позволяющую осуществлять речевой ввод и форматирование текста в текстовом процессоре MS Word. На практике эти программы показывают недостаточно высокие результаты (при тестировании точность не достигла даже 90% [1]).

Данная статья посвящена разработке системы речевого управления, которая основана на сегментно-целостной модели восприятия речевого сигнала. Эта бионическая модель базируется на

представлении о мозге как о двухканальной системе применительно к обработке звуковой речи [2]. Каналы сегментного и целостного восприятия, соответствующие левому и правому полушариям головного мозга, действуют параллельно, обеспечивая высокую скорость и надёжность распознавания. В статье предлагается реализация первого канала в виде нейросетевого пофонемного распознавателя, а второго – в виде нечёткого классификатора целостных паттернов (слов). Взаимодействие этих каналов для решения задачи речевого управления текстовым процессором MS Word показано на рис.1.

Целостный подход к распознаванию речи основан на сопоставлении паттернов входного сигнала с хранящимися в памяти целостными эталонами. В качестве паттернов рассматриваются слова, набор которых определяет словарный состав речевого командного интерфейса с редактором MS Word.

Речевое слово представляется в виде двумерного спектрального временного образа (СВО), получаемого с помощью оконного преобразования Фурье (рис.1а). СВО позволяет выделить местоположение резонансных частот, т.е. локальных выбросов, что является определяющей особенностью речевого сигнала [3]. На этом основании СВО можно преобразовать к двоичному виду с помощью замены: 1 – на месте локального выброса, 0 – в других местах. Полученный образ является двоичным спектральным временным образом (ДСВО) и используется как отражение особенностей речевого сигнала (рис.1б).

Для корректного сопоставления речевых образов необходимо провести их выравнивание по длине. Эта процедура выполнена с помощью нелинейного выравнивания, учитывающего неравномерность протекания сигнала во времени [4], для чего использовался алгоритм, основанный на определении наилучшего соответствия входных и эталонных речевых образов, известный как метод DTW [5]. В отличие от алгоритма линейного приведения длин, применяемый алгоритм осуществляет выравнивание входного ДСВО и эталонного образа только на спектрально подобных фрагментах.

Для распознавания изолированных слов, нормализованных по времени, применялся метод нечёткого сопоставления с эталоном [3]. Эталонные образы для каждого слова словаря формировались как среднее арифметическое ДСВО различных вариантов произношения этого слова. В результате формируется бинарное нечёткое отношение между множеством F (номеров частот f) и множеством T (номеров временных интервалов t) в виде:

$$f \in F,\ t \in T\ :\ F\ R\ T\ ,$$

где R – нечёткое отношение, которое ставит каждой паре элементов $(f,t) \in F \times T$ величину функции принадлежности $\mu_R(x,y) \in [0,1]$. Набор нечётких отношений $R = \{r_1, r_2, \ldots r_n\}$ определяет словарь эталонов размером n.

Распознаваемый образ y рассматривается как обычное (чёткое) отношение между множеством частот и множеством временных интервалов. Для него вычисляются степени сходства S_j с каждым нечётким отношением r_j , и результатом распознавания является номер j слова в словаре, такой, что

$$j = \max_{j \in [1,n]} \{S_j\},$$

где

$$S_j = \frac{\int r_j(f,t) \wedge y(f,t) dfdt}{\int \neg r_j(f,t) \wedge y(f,t) dfdt}.$$

Были проведены экспериментальные исследования, направленные на определение качества распознавания речевых команд по методу нечёткого сопоставления при линейном и нелинейном выравнивании образов. Для эксперимента использовалась речевая однодикторная база данных, включавшая в себя звукозаписи 6 речевых команд управления текстовым процессором: “Автоформат”, “Жирный”, “Курсив”, “Маркеры”, “Найти”, “Нумерация”. Каждая речевая команда была представлена 30 реализациями, 15 из которых использовались для обучения системы, а 15 – для тестирования. Результаты распознавания слов тестового множества представлены в табл. 1.

Таблица 1. Результаты тестирования системы

	Автоформат	Жирный	Курсив	Маркеры	Найти	Нумерация	**Итого, %**
Автоформат	15	0	0	0	0	0	**100,00**
Жирный	0	15	0	0	0	0	**100,00**
Курсив	0	0	15	0	0	0	**100,00**
Маркеры	0	0	0	15	0	0	**100,00**
Найти	0	0	0	0	15	0	**100,00**
Нумерация	0	0	0	0	0	15	**100,00**
Качество распознавания составило **100,00%**							

Сегментный подход к распознаванию речи основан на фонетическом анализе речевого сигнала. Предложен метод, основанный на определении меры сходства фрагмента речевого

сигнала с каждой из фонем с последующим выбором наиболее достоверной фонетической цепочки [6].

Пусть $A_w(t)$ – акустическое представление высказывания w; $F_k(t)$ – акустическое представление некоторой фонемы. Требуется определить, является ли фонема, описываемая $F_k(t)$, фрагментом высказывания $A_W(t)$.

Представим $F_k(t)$ на отрезке $[t_0, t_1]$ в виде множества пар

$$\{(X'(t), Y'(t))\}, \qquad (1)$$

где $\mathbf{X}'(t)=(F_k(t-m), F_k(t-m+1),\ldots F_k(t-1))$, $m=const$; $Y'(t)=F_k(t)$; $t_0 \le t \le t_1$. Аналогично представим $A_w(t)$ в виде множества пар $\{X(t), Y(t)\}$.

Представление $F_k(t)$ в виде (1) позволяет сформировать нейросетевую функцию NET: $NET(X'(t)) = Y'(t)$. Тогда мера отличия Err_k участка $A_w(t)$ при $t \in [t_н, t_к]$ от $F_k(t)$ определяется: $Err_k(t) = \left| Y(t) - NET(X(t)) \right|$.

Таким образом, получаем новое параметрическое описание исходного сигнала:

$$A_w(t) \rightarrow (Err_1(t), Err_2(t) \ldots Err_n(t)),$$

где $Err_k(t)$ – мера отличия участка сигнала $A_w(t)$ от k-й фонемы на фрагменте сигнала длительности m.

Новое параметрическое описание исходного сигнала имеет преимущества, связанные с более высокой стабильностью описания на стационарных участках, а также с интерпретируемостью полученных величин. Однако сложная форма и значительная нестабильность речевого сигнала не позволяют сделать вывод о фонеме по отдельным мгновенным значениям мер отличия $Err_k(t)$. Поэтому результаты распознавания усредняются на достаточно большом участке времени. Полученное параметрическое описание сигнала используется при дальнейшей контекстной обработке, как это показано на схеме распознавания (рис. 3).

Первый уровень схемы состоит из набора нейронных сетей, каждая из которых обучена распознаванию отдельной фонемы. Выходы сетей интерпретируются как прогноз следующих значений сигнала при условии, что имеет место соответствующая фонема. На втором уровне ошибка прогноза накапливается на всей протяженности окна сегмента речи. Интегральная ошибка поступает на третий уровень, где из всех фонем выбираются наилучшие. Полученный набор участвует в формировании фонетических цепочек, представляющих собой гипотезы о произносимом слове.

Произнесённое слово определяется по цепочке с наибольшей степенью достоверности.

Работа метода проиллюстрирована на примере распознавания слова «один» (рис.4). В примере задействовано четыре фонемы. Для аппроксимации использованы трехслойные сети типа «многослойный персептрон» с 20 входами и количеством нейронов в слоях 20-10-1. Количество входов определялось в соответствии с оценкой периода основного тона для данного диктора. В обучении использовано по 10 реализаций каждой фонемы одного диктора.

На рис.4 видно, что минимумы ошибки последовательно достигаются на участках сигнала, соответствующих заданным фонемам.

На основе метода был разработан модуль распознавателя, словарь которого включал 60 речевых команд. Точность распознавания этих команд превысила 90%.

Предложенные способы речевого управления текстовым редактором были реализованы в виде программного компонента ActiveX, интегрируемого с системой MS Word. В результате тестирования найден оптимальный баланс по критерию эргономичности между речевой и тактильно-зрительной составляющей интерфейса с текстовым редактором. Через речевой канал интерфейса целесообразно организовывать передачу наиболее употребительных команд, а также макрокоманд (последовательностей простых действий), связанных со сложным редактированием документа. Тактильно-зрительный канал целесообразно использовать для передачи команд, связанных с позиционированием фрагментов документа в пространстве.

1. Программы синтеза и распознавания речи. Тестовая лаборатория. – http://art.bdk.com.ru/govor/1listr62t.htm.

2. Восприятие речи: вопросы функциональной асимметрии мозга / Морозов В.П., Вартанян И.А., Галунов В.И. и др. – Л.: Наука, 1988. – 135 с.

3. Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др. Распознавание речи // Прикладные нечёткие системы. Под редакцией Т.Тэрано, К. Асаи, М. Сугено. – М.: “Мир”, – 1993. – 157-170 с.

4. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – Киев: Наукова думка. – 1987. – 264 с.

5. Stuart N. Wrigley. Speech Recognition by Dynamic Time Warping. – http://www.dcs.shef.ac.uk/~stu/com326/index.html.

6. Федяев О.И., Гладунов С.А. Нейросетевой интерпретатор речевых команд для управления программными системами. – Труды 7-й всероссийской конференции “Нейрокомпьютеры и их применение”, НКП-2001, Москва, 14-16 февраля 2001 г./ Под редакцией А.И. Галушкина. М.: Институт проблем управления, 2001 – с. 298-301.

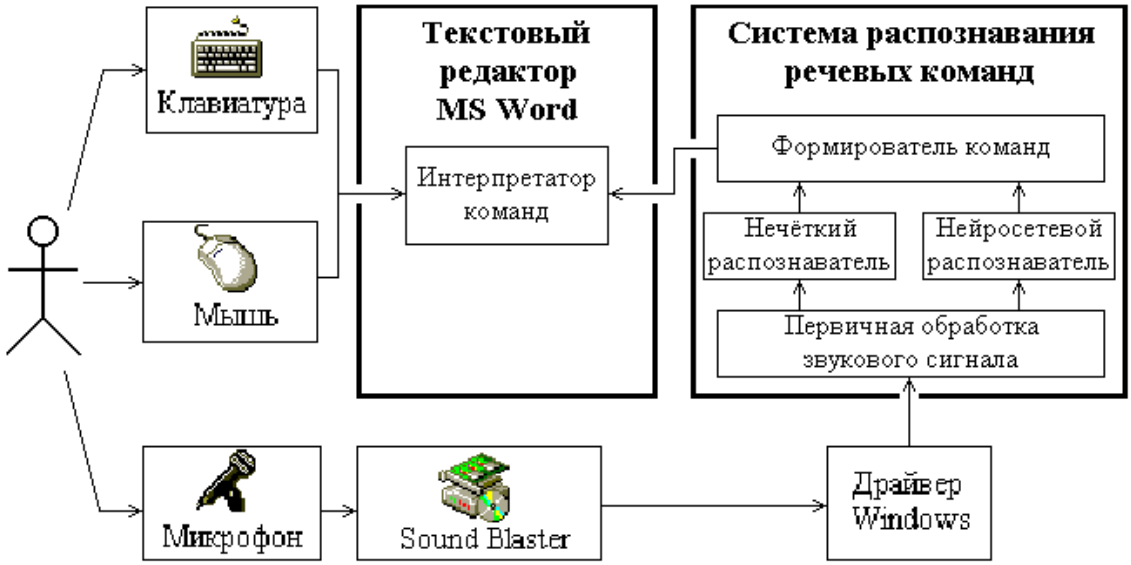

Рисунок 1 – Структура речевого канала управления текстовым редактором MS Word

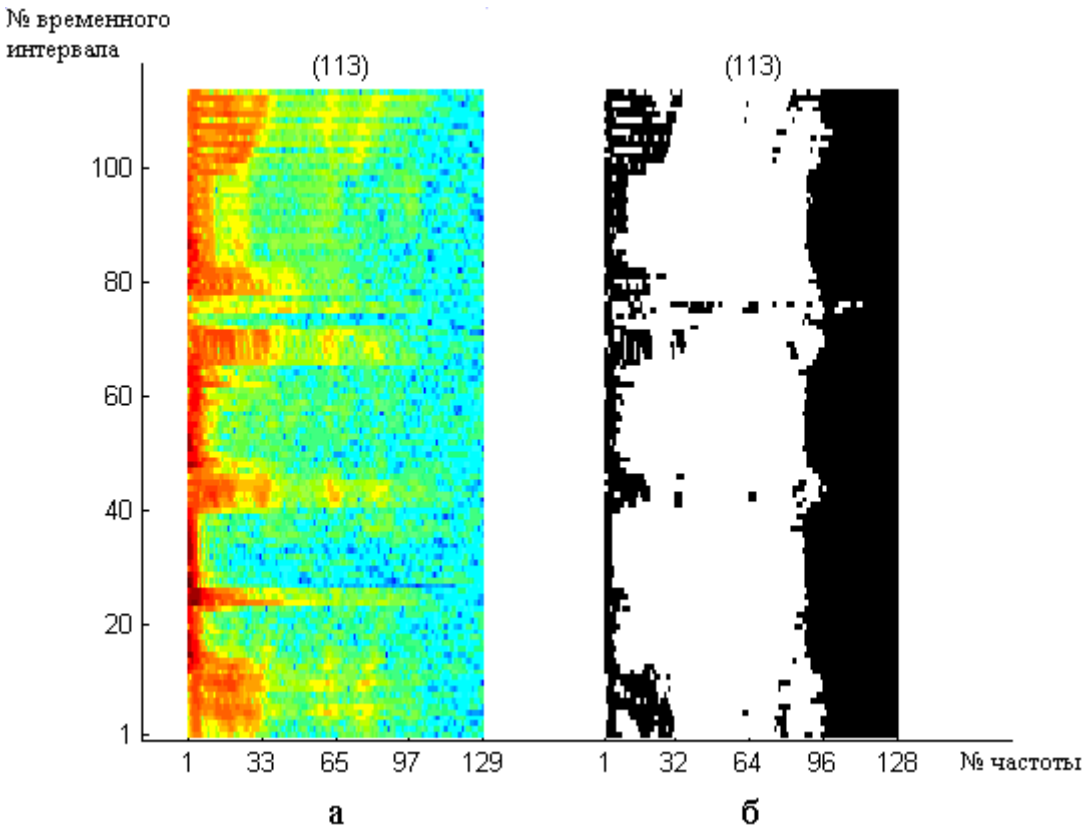

Рисунок 2 – Пример спектрально-временного представления слова “автоформат”: а – СВО; б – ДСВО

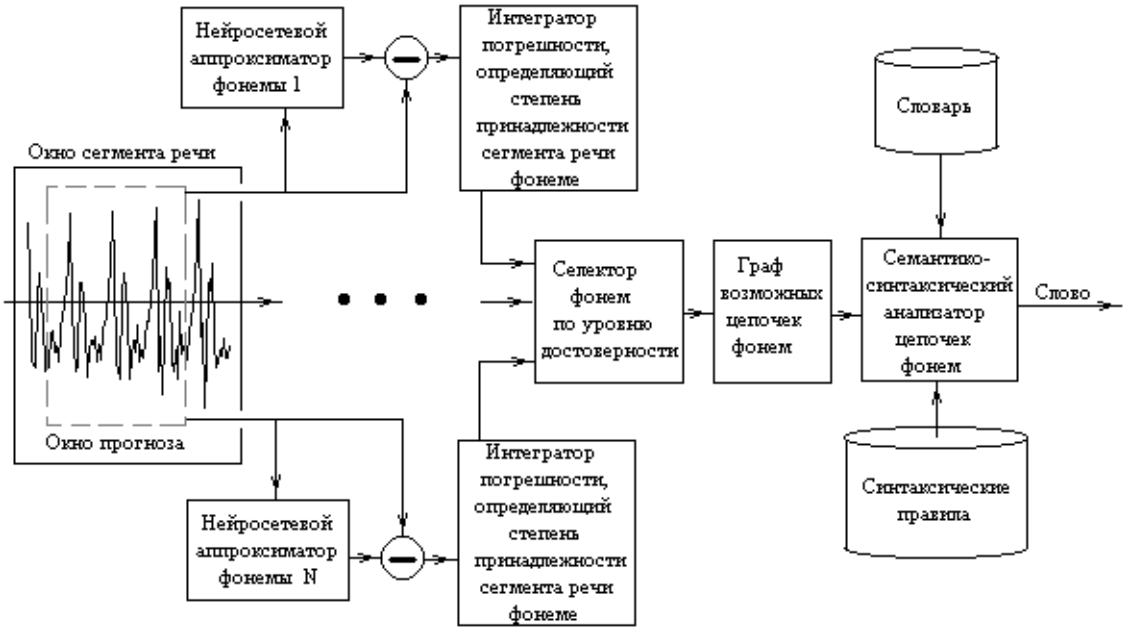

Рисунок 3 – Схема распознавания речи на основе нейросетевой аппроксимации фонем

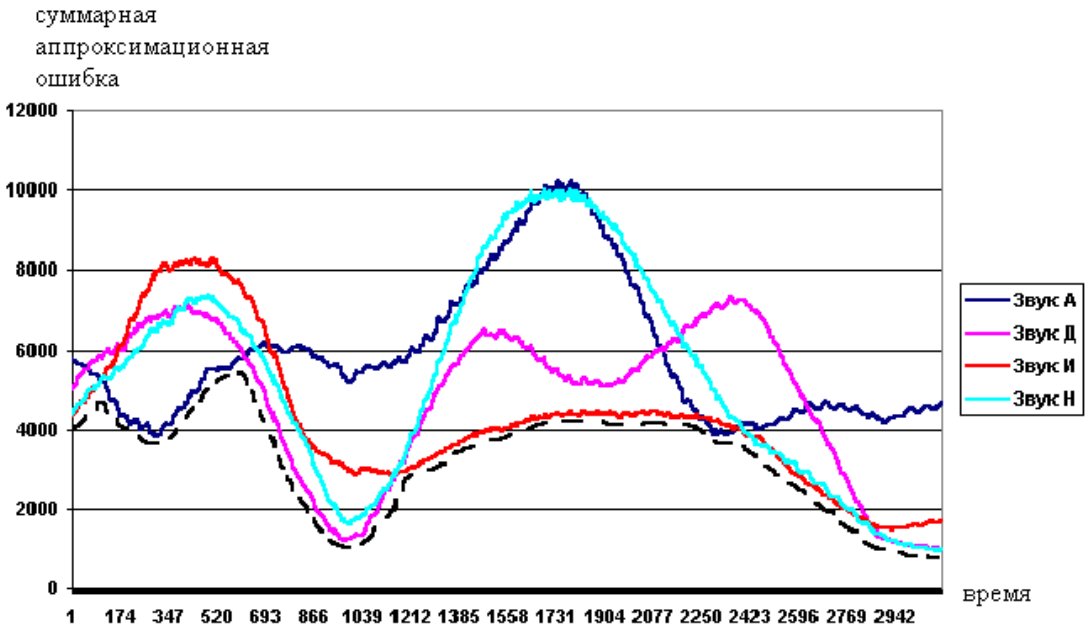

Рисунок 4 – Графики меры отличия входного сигнала от различных фонем для двух вариантов произнесения слова «один»: (- - -) – линия наименьших ошибок, определяющая наилучшую цепочку фонем