СОГЛАСОВАНИЕ РЕШЕНИЙ СЕГМЕНТНОГО И ЦЕЛОСТНОГО КАНАЛОВ В ДВУХКАНАЛЬНОЙ МОДЕЛИ РЕЧЕВОГО УПРАВЛЕНИЯ

Федяев О.И., Гладунов С.А., Бондаренко И.Ю.
Кафедра ПМиИ ДонНТУ
fedyaev@r5.dgtu.donetsk.ua

Abstract

A speech control module structure is considered which is based on a co-operation of segment and integral speech recognition channels. The first channel is implemented in basis of neural mathematics, and the second one represents fuzzy computations. An approach of two channels decisions coordination is proposed. Solution is based on certainty factors method. Applying this method allows generalizing solution of recognition channels as independent experts and work out a more precise decision about a word recognized.

Введение

В связи с расширением сферы использования средств вычислительной техники и их усложнением возникает необходимость разработки более простых и удобных для людей способов взаимодействия человека и компьютерных систем. Одним из таких способов является речевой человеко-машинный интерфейс, поскольку речь – это наиболее естественное для человека средство обмена информацией. О необходимости речевого интерфейса свидетельствует и возросшее число коммерческих разработок систем, использующих такой интерфейс. Например, программная система 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]). Таким образом, задача разработки эффективного метода распознавания речи, составляющего основу речевого взаимодействия, по-прежнему является актуальной.

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 работе [3] предложена общая структура сегментно-целостной системы речевого управления, а также реализация сегментного и целостного каналов, но не рассмотрена задача формирования коллективного решения на основе интеграции результатов работы каналов. Поэтому предметом данной работы являются методы согласования решений сегментного и целостного каналов в двухканальной модели речевого управления.

1. Структура двухканальной системы распознавания речевых команд

В основу двухканальной системы речевого управления положены современные представления о механизмах речевой деятельности человека [2]. Структурная схема работы двухканальной системы распознавания речевых слов представлена на рис. 1.

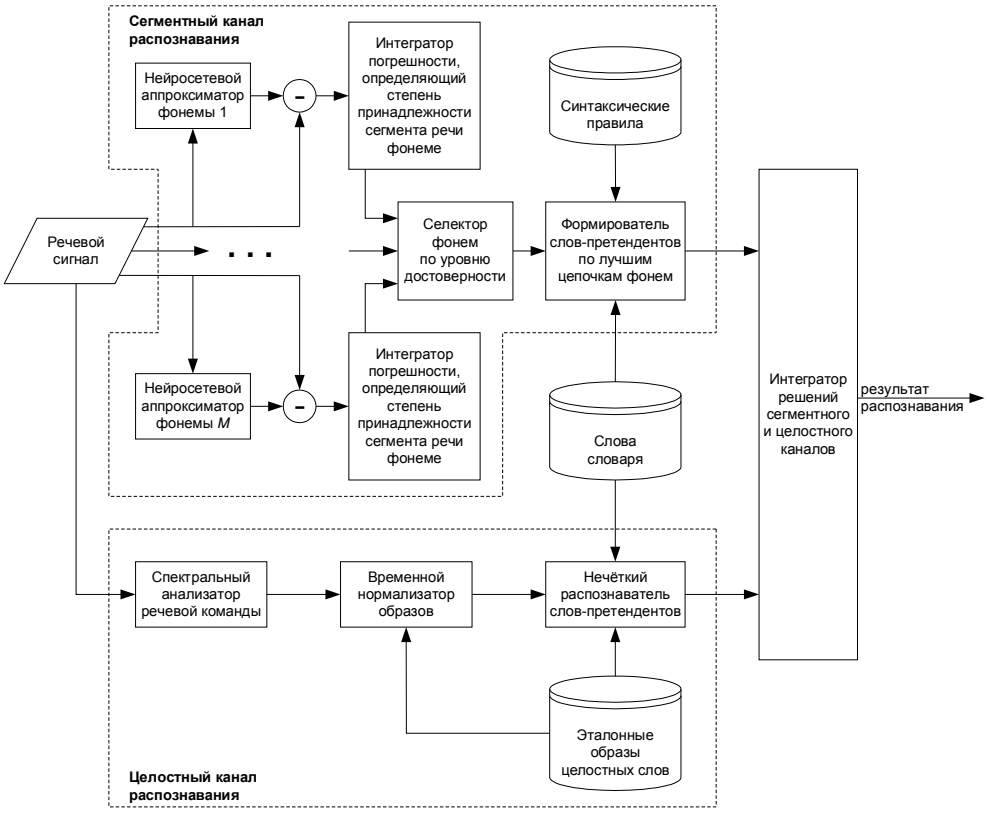

Рис. 1. Структурная схема двухканальной модели распознавания речи

В звуковом сигнале, поступающем на вход системы, определяются границы речевого участка – предполагаемой речевой команды – на основе функций кратковременной энергии сигнала, числа переходов через нуль и

количества точек постоянства. Далее выделенная речевая команда параллельно анализируется сегментным и целостным каналами. Сегментный канал основан на методе скользящего фонетического анализа [4], а целостный канал – на методе нечёткого DTW-сопоставления образов [5]. Эти каналы формируют независимые наборы слов-претендентов, т.е. слов, к каждому из которых с определённым коэффициентом уверенности может быть отнесена распознаваемая команда.

На последнем уровне системы, используя наборы слов-претендентов и соответствующие им коэффициенты уверенности, проводится согласование приближённых решений сегментного и целостного каналов и принимается окончательное решение о распознаваемой команде.

2. Целостный канал распознавания

Для разработки целостного канала распознавания предложен метод нечёткого DTW-сопоставления образов [5]. По сравнению с методом нечёткого сопоставления образов [6] DTW-сопоставление, использующее алгоритм Dynamic Time Warping (DTW) для нелинейной временной нормализации сравниваемых образов, позволяет повысить качество распознавания речевых команд.

В качестве единиц речи рассматриваются слова, набор которых определяет словарный состав речевого командного интерфейса. Речевой сигнал представляется в виде двоичного двумерного спектрального временного образа (ДСВО), позволяющего выделить местоположение резонансных частот, т.е. локальных выбросов, что является определяющей особенностью речевого сигнала [6].

Временную нормализацию речевых образов предложено выполнять с помощью нелинейного выравнивания, учитывающего, в отличие от простого линейного выравнивания, неравномерность протекания сигнала во времени [7]. В основу алгоритма нелинейного выравнивания был положен метод DTW.

Для распознавания изолированных слов, нормализованных по времени, применялся метод нечёткого сопоставления с эталоном [6]. 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 в соответствие каждой паре элементов $(f,t) \in F \times T$ значение 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 в качестве результата распознавания принимается такой номер j слова в словаре, что $j = \max\limits_{j\in[1,n]} \{S_j\}$, где

$$S_j = \int r_j(f,t) \wedge y(f,t) df dt \cdot \left(\int \overline{r_j(f,t)} \wedge y(f,t) df dt \right)^{-1}.$$

3. Сегментный канал распознавания

Сегментный подход к распознаванию речи основан на фонетическом анализе речевого сигнала. Использован метод нейросетевой аппроксимации фонем, основанный на определении меры сходства фрагмента речевого сигнала с каждой из фонем с последующим выбором наиболее достоверной фонетической цепочки [4]. Метод позволяет с некоторой погрешностью установить, является ли фонема, описываемая $F_k(t)$, фрагментом высказывания $A_W(t)$, где $A_w(t)$ – акустическое представление высказывания w; $F_k(t)$ – акустическое представление некоторой фонемы. С этой целью функция $F_k(t)$ на отрезке $[t_0 , t_1]$ представляется в виде множества пар

$$\{(X'(t), Y'(t))\}, \qquad (3.1)$$

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

Представление $F_k(t)$ в виде (3.1) 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 по формуле

$$Err_k(t) = \left| Y(t) - NET(X(t)) \right|.$$

Таким образом, формируется 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

усреднялись на достаточно большом участке времени. Полученное параметрическое описание сигнала используется при дальнейшей контекстной обработке, как это показано на схеме распознавания (рис. 1).

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.

4. Интегратор решений сегментного и целостного каналов

На последнем этапе работы системы распознавания возникает необходимость в формировании коллективного решения на основе интеграции результатов работы нечёткого и нейросетевого каналов. С этой целью вводится интегратор, формирующий итоговое решение в условиях ненадёжной информации. Учитывая неточность выводов каждого из каналов, для принятия итогового решения применялись методы коллективного распознавания [8]. Одним из первых подходов к принятию решений в среде ненадёжных знаний был подход, основанный на байесовской вероятности [9]. Вследствие невозможности осуществления корректных вероятностных рассуждений, позже был предложен метод коэффициентов уверенности [10], ставший классическим примером обработки ненадёжных знаний. Известны и другие способы принятия решений в среде ненадёжных знаний: метод Демпстера-Шафера [10, 11], методы нечёткой логики [11], субъективный байесовский метод [10], нейросетевые методы [11], метод голосования [8].

Блок интегратора реализует метод неточных рассуждений на основе фактора уверенности, предложенный для системы MYCIN [10]. Выбор метода обусловлен его простотой и хорошей способностью к компромиссу при объединении мнений независимых экспертов (каналов распознавания). Блок интегратора, представленный на рис.2, осуществляет прямой логический вывод на основе утверждений каналов распознавания с коэффициентами уверенности *CF* (Certainty Factor) и знаний о формировании коллективного решения группы независимых экспертов.

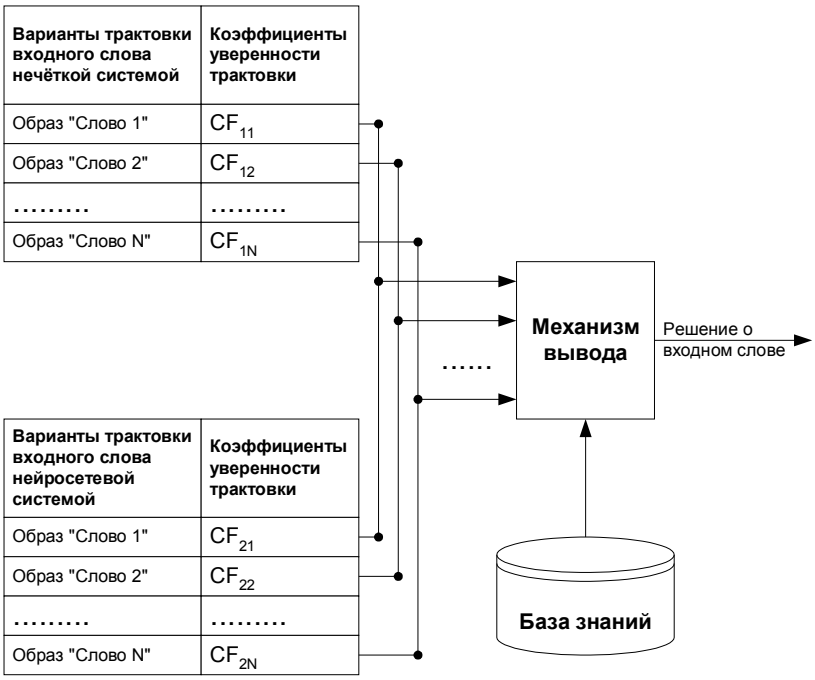

Рис. 2. Экспертное заключение по решениям каналов распознавания

В задачах коллективного распознавания с ненадёжными данными важную роль играет комбинированная связь, обозначаемая как КОМБ [10]. Она независимо подкрепляет или опровергает выдвигаемую гипотезу о произнесённом слове на основании двух и более мнений экспертов. Поэтому знания, представленные в виде продукционных правил (рис. 3), предусматривают комбинированную связь между решениями, предлагаемыми каналами распознавания. Допустим, что один из экспертов каждого канала уже определил степени надёжности X и Y как результат предварительного распознавания, и необходимо сделать вывод (вычислить степень надёжности заключения V), используя правила из базы знаний интегратора решений. Прямой логический вывод итогового заключения основан на использовании известной методики оценки предпосылки правил (антецедента) и учёта связи КОМБ по методу MYCIN. Степень надёжности распространяется по иерархической сети логического вывода (рис. 4), образуемой продукционными правилами.

В методе MYCIN ненадёжность фактов представляется коэффициентом уверенности CF, принимающим значения от +1 (если факт заведомо истинный) до 0 (для заведомо ложных фактов). Запись $CF[V, Z]$ будем трактовать как коэффициент уверенности в истинности вывода V, если удовлетворяется предпосылка Z (консеквент). В процессе вывода при наличии связи КОМБ отдельно вычисляются $CF[V_i , X_i]$ и $CF[V_i , Y_i]$ по формуле:

$$CF[V, Z] = CF_{\text{правила}} \cdot CF_{\text{предпосылки}} ,$$

где $Z \in \{X, Y\}$; $CF_{\text{правила}} = 1$; $CF_{\text{предпосылки}} = CF_{ij}$.

Объединение решений экспертов по соответствующим словам словаря осуществляется с помощью комбинированной функции

$$CF[V_i(X_i, Y_i)] = CF[V_i, X_i] + CF[V_i, Y_i] - CF[V_i, X_i] \cdot CF[V_i, Y_i].$$

Результат распознавания (номер распознанного слова) формируется как номер максимального элемента в массиве $CF[V_i]$ (рис. 4).

1. ЕСЛИ Трактовка нечёткой системой входного образа = образ «Слово 1» [CF_{11}]

   ТО Версия 1 = «Слово 1»

2. ЕСЛИ Трактовка нечёткой системой входного образа = образ «Слово 2» [CF_{12}]

   ТО Версия 2 = «Слово 2»

. . . . . .

8. ЕСЛИ Трактовка нейросетевой системой входного образа = образ «Слово 1» [CF_{21}]

   ТО Версия 1 = «Слово 1»

9. ЕСЛИ Трактовка нейросетевой системой входного образа = образ «Слово 2» [CF_{22}]

   ТО Версия 2 = «Слово 2»

. . . . . .

Рис. 3. База знаний

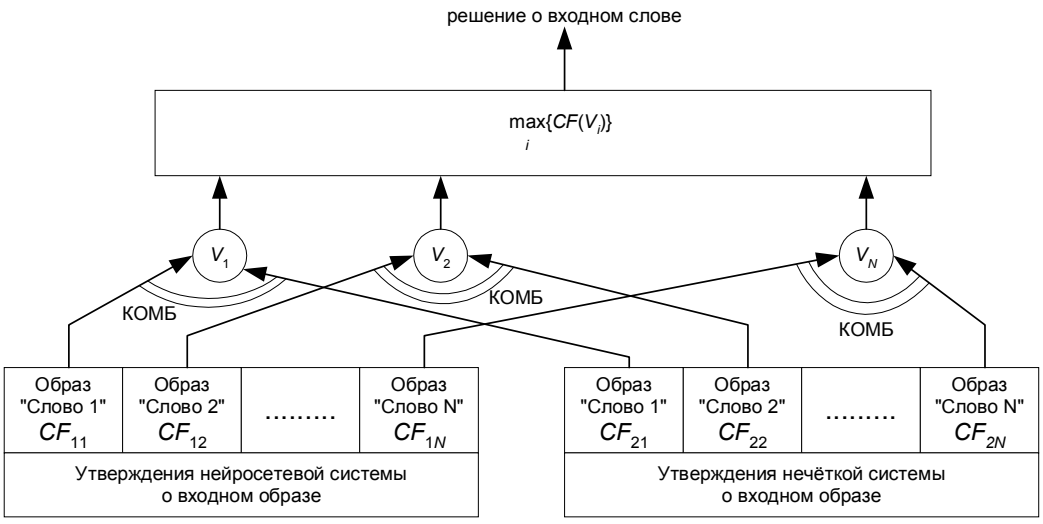

Рис. 4. Схема вывода с учётом ненадёжности знаний

На рис. 5 приведены результаты коллективного распознавания слова “Маркеры” в сопоставлении с другими словами словаря. В данном примере сегментный (нейросетевой) канал явно не отдаёт предпочтения какому-то одному слову, а с учётом мнения другого эксперта – целостного (нечёткого) канала – интегратор вырабатывает коллективное решение, которое более уверенно идентифицирует входное слово как “Маркеры”.

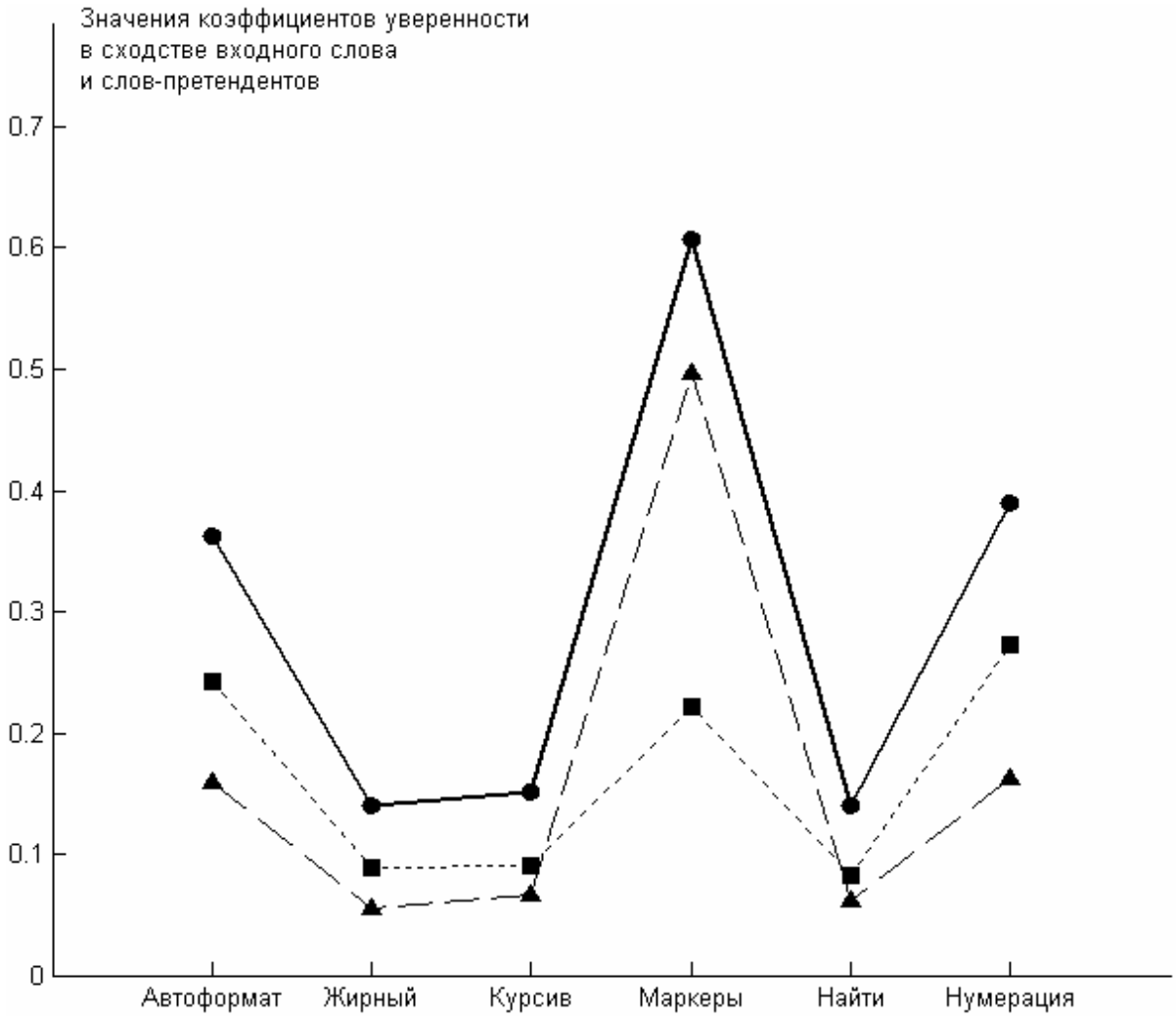

Рис. 5. Результаты коллективного распознавания слова “Маркеры”:
(▲— —▲)– значения коэффициентов уверенности целостного канала;
(■------■)– значения коэффициентов уверенности сегментного канала;
(●——●)– значения коэффициентов уверенности после интеграции.

Заключение

Предложен способ согласования решений сегментного и целостного каналов в двухканальной модели речевого управления, основанный на методе коэффициентов уверенности. Применение этого метода позволило учесть коллективное мнение каналов распознавания как

независимых экспертов и принять более точное решение о распознаваемом слове. Объединение решений каналов осуществляется интегратором, представляющим собой экспертную систему, которая выполняет логический вывод решения на основе ненадёжных знаний. Эксперименты показали, что данный способ позволяет эффективно разрешать конфликтные ситуации, возникающие при расхождении мнений экспертов (каналов распознавания).

Перечень ссылок

1. Информационный портал речевых технологий “Голосовые технологии”. Тестовая лаборатория. – 6.04.2006. – http://art.bdk.com.ru /govor/1listr62.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. Бондаренко И.Ю., Гладунов С.А., Федяев О.И. Сегментно-целостная структура канала речевого управления программными системами // Сб. трудов X нац. конференции по искусств. интеллекту с междунар. участием КИИ-2006. – М.: Физматлит, 2006. – с. 841 – 849.
4. Гладунов С.А. Аппаратно-программные средства раздельной локализации фонем в системах речевого взаимодействия человека с ЭВМ: Автореф.дис…канд.техн.наук: 05.13.13 / ДонНТУ. – Донецк, 2005. – 22 с.
5. Бондаренко И.Ю., Федяев О.И. Анализ эффективности метода нечёткого сопоставления образов для распознавания изолированных слов // Сб. трудов VI междунар. науч.конференции “Интеллектуальный анализ информации ИАИ-2006”. Под ред. Таран Т.А. – К.: Просвіта, 2006. – с.20–27.
6. Киедзи Асаи, Дзюндзо Ватада, Сокуке Иваи и др. Распознавание речи // Прикладные нечёткие системы: Пер. с яп. Под ред. Т.Тэрано, К. Асаи, М. Сугено. – М.: Мир, 1993. – с. 157-170.
7. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. – К.: Наукова думка, 1987. – 264 с.
8. Городецкий В.И., Серебряков С.В. Методы и алгоритмы коллективного распознавания: Обзор // Труды СПИИРАН, том 1, вып. 3. – С.-Пб.: Наука, 2006. –139 – 171 с.
9. Джексон П. Введение в экспертные системы. – М.: “Вильямс”, 2001. – 624 с.
10. Представление и использование знаний: Пер. с япон./ Под ред. Х.Уэно, М.Исидзука. – М.: Мир, 1989. – 220с.
11. Люгер Д.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. – М.: “Вильямс”, 2003. – 864 с.

Дата надходження до редакції 25.10.2006 р.