УДК 004.934

Т.В. Шарий
Донецкий национальный университет, Украина
Украина, 83001, г. Донецк, ул. Университетская, 24

Голосовое управление мобильным роботом на основе когнитивной модели FCAS

T.V. Sharii
Donetsk National University, Ukraine
Ukraine, 83001, c. Donetsk, 24 Universitetskaya str.

A Mobile Robot Voice Control Based on the Fcas Cognitive Model

Т.В. Шарій
Донецький національний університет, Україна
Україна, 83001, м. Донецьк, вул. Університетська, 24

Голосове керування мобільним роботом на основі когнітивної моделі FCAS

В статье рассматривается задача дикторонезависимого распознавания команд голосового управления роботом в реальных условиях. На этапе постобработки речевого сигнала используется легковесная нечеткая когнитивная модель FCAS. Экспериментально обосновано применение данной модели.
Ключевые слова: FCAS, когнитивная модель, фонологический признак, просодия.

The article deals with the task of speaker-independent voice control command recognition under the real conditions. The lightweight fuzzy cognitive model FCAS is used at the stage of a speech signal post-processing. The use of the model is experimentally grounded.
Key words: FCAS, cognitive model, phonological feature, prosody.

Розглядається задача дикторонезалежного розпізнавання команд голосового керування роботом у реальних умовах. На етапі постобробки мовленнєвого сигналу використовується легковага нечітка когнітивна модель FCAS. Експериментально обґрунтовано застосування даної моделі.
Ключові слова: FCAS, когнітивна модель, фонологічна ознака, просодія.

В настоящее время одним из приоритетных направлений искусственного интеллекта является социальная робототехника, в рамках которой обеспечивается взаимодействие человека с автономными роботами-помощниками [1], [2]. Голосовое общение является одним из важнейших естественных способов диалога человека с роботом. В целом, за последние пять лет удалось достичь значительного прогресса в решении задачи компьютерного распознавания речи, а лидерами в данной области являются разработки GoogleSpeech от Google [3] и Siri от Apple [4]. Эти технологии основываются на статистических моделях: комбинациях Скрытых марковских моделей (СММ) и нейронных сетей [5]. На этапе предобработки речевого сигнала используются, в основном, кепстральные коэффициенты на шкале мел (MFCC), позволяющие компактно описать спектр сигнала. Однако недостатком указанных программных решений является то, что они требуют связи с интернетом и доступа к большим коллекциям речевых данных для обучения. Кроме того, мобильным роботам необходимо учитывать специфический шум от собственного движения. В связи с этим, в робототехнике используются иные аппаратно-программные решения, например [6], [7], базирующиеся на

ограниченных словарях, включающих, в среднем, 40 – 50 команд. В целом, автоматическое распознавание речи в робототехнике остается нетривиальной проблемой. В робототехническом распознавании речи выделяют две задачи: локализация речевого потока и распознавание его содержания.

Целью данной статьи является решение второй задачи в условиях реальных помещений для мобильного робота с ограниченным словарем. Специфическим требованием к решению является его некритичность к ресурсам для обеспечения функционирования модуля в режиме реального времени. В условиях шума помещений и использования микрофонов относительно низкого качества важно автоматически выделять наиболее информативные участки речевого сигнала.

Основная идея статьи заключается в применении нечеткой когнитивной модели постобработки речевых сигналов FCAS (Fuzzy Cognitive Accented Speech) [8] для дикторонезависимого распознавания голосовых команд управления роботом. Данная модель является относительно легковесной и не требует тренировки на произношение конкретного диктора. Ее удобство заключается также в возможности настройки свободных параметров для конкретных ситуаций и приложений.

Общая структура модели FCAS. Главным блоком модели FCAS является ядро, в котором обрабатывается нечеткая информация, получаемая от двух блоков – блока расчета весов речевых сегментов и признакового (фонологического) блока (рис. 1). В общем случае модель также содержит блок фонемного анализа [8], однако в задаче распознавания ограниченного набора голосовых команд его можно исключить из общей схемы с целью поддержания легковесности системы; при этом функции фонемного блока принимает на себя признаковый блок, о чем будет сказано далее.

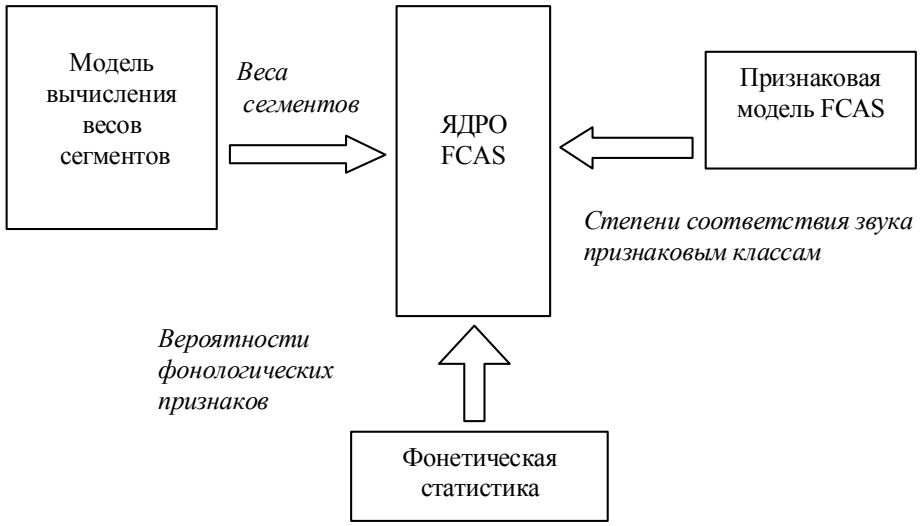

Рисунок 1 – Схема модели FCAS

Основным отличием предлагаемой модели от традиционно используемых в современных речевых технологиях СММ является многоуровневый учет акустико-фонетической информации на этапе постобработки речевого сигнала. В модели FCAS не только закладывается статистика звуков речи и слов любого языка с их спектральными и кепстральными прототипами, но и осуществляется принятие решений на основе такой информации, как: вес сегмента (модель вычисления весов FCAS), фонологический класс сегмента (признаковая модель FCAS) и последовательность сменяющихся сегментов звуков речи (ядро FCAS).

Взвешивание речевых сегментов в модели FCAS. Восприятие человеком фонем, слогов и слов в значительной степени определяется наиболее «ярко» звучащими фрагментами речи. При этом на субъективное восприятие яркости звучания, определяющее влияние, оказывают просодические признаки (такие, как частота основного тона, энергия и длительность) [9]. В работе исследовались для анализа просодии дескрипторы, применяющиеся в задачах анализа музыки [10]: темпоральные (описывающие картину временны*х* изменений сигнала в процессе звучания), гармонические (основываются на гармонической структуре определенных звуков, в том числе гласных звуков речи) и перцепционные (основываются на особенностях восприятия звуков человеком). В качестве основного был выбран гармонический дескриптор «Разгармонизация» [10], характеризующий расхождение спектра данного звука со спектром гармонического звука:

$$INH = \frac{2}{f_0}\frac{\sum_h |f(h) - h \cdot f_0| \cdot a^2(h \cdot f_0)}{\sum_h a^2(h \cdot f_0)}, \quad (1)$$

где $a(f)$ – амплитудный спектр звука; f_0 – частота основного тона; h – номер гармоники основного тона (общее число анализируемых гармоник равно 12); $f(h)$ – «реальная» h-ая гармоника (которой соответствует пик в спектре). Вокализованные звуки имеют малое значение INH, а шипящие звуки – большое. Данные два класса звуков важны с точки зрения восприятия человеком, в связи с чем для вычисления веса сегмента используется следующая формула:

$$w = \max\{AM, INH\}, \quad (2)$$

где INH – разгармонизация сегмента (1), AM – степень ударности сегмента, которая рассчитывается по формуле:

$$AM = k_L L_r + k_E E_r + k_F \Delta f_0, \quad (3)$$

где k_L – коэффициент, характеризующий влияние относительной длительности сегмента L_r на его ударность; k_E – коэффициент влияния относительной энергии E_r на ударность; k_F – коэффициент влияния изменения частоты основного тона на ударность.
В работе экспериментально были подобраны следующие значения: $k_L = k_E = 0.3$, $k_F = 0.4$.

Показатель изменения частоты основного тона Δf_0 рассчитывается по формуле:

$$\Delta f_0 = \begin{cases} 1 & ,\delta > 1 \\ \dfrac{\delta + 1}{2} & ,\delta \in [-1, 1], \\ 0 & ,\delta < -1 \end{cases} \quad (4)$$

где δ – среднее изменение гармоник основного тона:

$$\delta = \frac{1}{n}\sum_{h=1}^{n}\frac{f_2(h) - f_1(h)}{h \cdot f_0}, \quad (5)$$

где $f_1(h)$ и $f_2(h)$ – h-ые гармоники основного тона в начале и конце сегмента соответственно; n=12 – количество анализируемых гармоник.

Нечеткая модель фонологической классификации FCAS. В признаковом блоке FCAS на основе результатов спектрального анализа речевого сигнала происходит вычисление значений функции принадлежности текущего звука речи признаковым классам бинарной схемы Вайрена-Штубса («Шипящий», «Высокий», «Диффузный» и т.д.) [11]. Бинарная система селекции дифференциальных признаков звука речи подвержена

ошибкам, поэтому в работе применяется ее нечеткий аналог (рис. 2). Для нечеткой классификации звуков речи используются числовые величины, отражающие акустические свойства выраженности фонологического признака, а также применяются процедуры фаззификации для каждого признакового класса на основе данных величин.

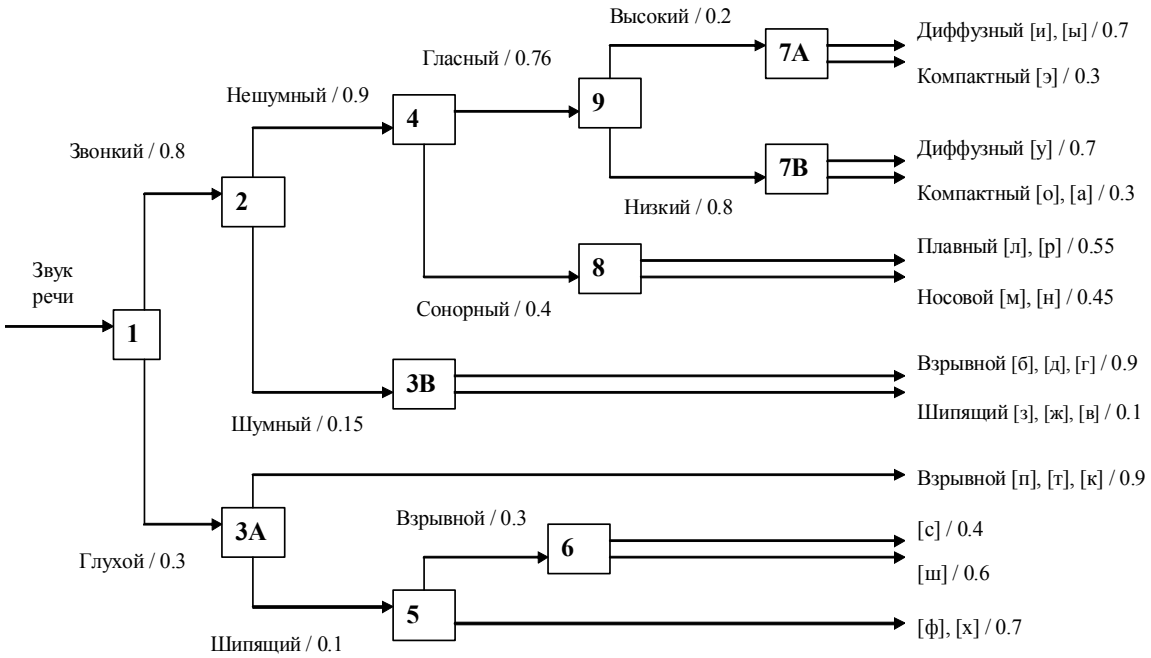

Рисунок 2 – Схема нечеткой фонологической классификации звуков речи

Для систем с ограниченными словарями необязательно учитывать все признаки, приведенные на рис. 2. Далее будет показано, что хороших результатов распознавания команд можно добиться, опираясь на четыре основных признака: компактный / диффузный, высокий/низкий, шипящий/взрывной, носовой/плавный.

Опишем, как вычисляются значения функции принадлежности звука речи данным признаковым классам.

В блоках 3А и 3В на рис. 2 происходит разделение звуков на взрывные и шипящие. В оригинальной схеме Вайрена-Штубса это разделение производится по уровню энергии сигнала на начальном участке звука [11]. Однако выделить автоматическим образом начальный участок произвольного звука довольно проблематично, а чаще всего невозможно (перед началом анализа система не имеет никакой информации о том, какая именно фонема звучит). С другой стороны, шипящие (сибилянтные) звуки имеют легко узнаваемый вид спектра – наличие шума в довольно широком диапазоне высокочастотной области. В связи с этим, в работе применяется степень сибилянтности звука (SBM, Sibilance Measure):

$$SBM = 1 - \frac{\sum_{i=400}^{1000} a(i) \cdot w(i)}{\sum_{j=1000}^{8000} a(j) \cdot w(j)}, \qquad (6)$$

где *a(f)* – амплитудный спектр участка звука; *w(f)* – весовая функция (значения весов для разных частотных диапазонов спектра берутся из табл. 1).

Таблица 1 – Весовые коэффициенты

Частотные диапазоны	100-700 Гц	700-1500 Гц	1500-2500 Гц	2500-4000 Гц	4000-7000 Гц	> 7000 Гц
Весовые коэффициенты	1	1.4	1.8	2.0	2.3	2.5

На основе характеристики *SBM* строится нечеткое множество «Сибилянтный»:

$$\underset{\sim}{SBM} = \{ SBM, \mu_{\underset{\sim}{SBM}}(SBM) \}, \qquad \mu_{\underset{\sim}{SBM}}(x) = \begin{cases} 0 & , x \le 0 \\ x & , 0 < x \le 1 \\ 1 & , x > 1 \end{cases} \tag{7}$$

Работа блоков 7A, 7B, 8 и 9 на рис. 2 также основана на принципе сравнения амплитуд в различных областях частотного спектра. В качестве числовых характеристик, отражающих акустические особенности каждого из этих признаков, в работе применяются степень диффузности (DM, Diffuseness Measure), степень высотности (ACM, Acuteness Measure) и степень назальности (NSM, Nasality Measure). Все предлагаемые характеристики имеют общий вид:

$$M(FNL, FNH, FDL, FDH) = \frac{\sum_{i=FNL}^{FNH} a(i) \cdot w(i)}{\sum_{j=FDL}^{FDH} a(j) \cdot w(j)}, \tag{8}$$

где *a(f)* – амплитудный спектр участка звука; *w(f)* – весовая функция (значения весов для разных частотных диапазонов спектра берутся из табл. 1); *FNL*, *FNH*, *FDL*, *FDH* – параметры, которыми отличаются указанные характеристики, и представляющие собой границы анализируемых частотных диапазонов спектра.

Значения *FNL*, *FNH*, *FDL*, *FDH* приведены в табл. 2.

Таблица 2 – Границы анализируемых частотных диапазонов спектра

Величина	FNL (Гц)	FNH (Гц)	FDL (Гц)	FDH (Гц)
DM	220	380	340	1100
ACM	2400	3700	800	1200
NSM	220	340	450	550

На основе степеней DM, ACM и NSM вводятся нечеткие множества «Диффузный», «Высокий» и «Носовой», соответственно. Каждое из них имеет S-образную функцию принадлежности вида, аналогичного (7).

Ядро FCAS. Рассчитанные для каждого звука весовые коэффициенты (2), а также значения функций принадлежности признаковым классам (6) – (8) используются в ядре FCAS, представляющем собой сеть взаимосвязанных элементарных фонетических процессоров (ЭФП) признакового и словесного уровней [8]. ЭФП каждого слова (команды) аккумулируют выходные значения признаковых ЭФП, с которыми они связаны. Например, на активность ЭФП слова «ИЩИ» будут влиять признаковые ЭФП классов «Высокий», «Диффузный» (звук [и]) и «Шипящий» (звук [щ]). Распознанным полагается слово, ЭФП которого имеет максимальное выходное значение. Вся сеть узлов названа «когнитивной», т.к. активация элементов, отвечающих за восприятие, производимая входным речевым сигналом, оставляет за собой «след» анализа

входа на каждом из уровней обработки (фонемном, признаковом, словесном). С этой точки зрения данная модель концептуально напоминает модель TRACE, предложенную Макклелландом и Элманом в [12]. Однако, в отличие от модели FCAS, в модели TRACE отсутствует учет весов речевых сегментов, и в ней фиксированы фонемные интервалы, что накладывает ограничения на ее практическое использование.

Вычисления FCAS происходят с периодичностью оконного анализа речевого сигнала (20 мс). Сигнал подвергается автоматической фонемной сегментации, в связи с чем в нем присутствуют маркеры фонетических сегментов (границы звуковых образов). Моменты времени, соответствующие данным границам, будем называть моментами принятия фонемного решения. В эти временные срезы модель выдает степени соответствия звучавшего речевого сегмента всем фонологическим признакам. Отметим, что для любого алгоритма сегментации характерен определенный процент лишних (вставляемых) границ сегментов. С учетом механизмов накопления и убавления состояний ЭФП эта ситуация «сглаживается», т.к. соседние сегменты будут соответствовать одному признаку (с различными лишь, возможно, степенями соответствия). Кроме этого, модель FCAS может выдать большую степень соответствия звуковому образу некоторой фонемы не только в момент принятия фонемного решения, но и в любой другой момент времени, что тоже будет обработано моделью. Таким образом, частично учитывается вторая проблема алгоритмов автоматической сегментации речевых сигналов – пропуск маркера сегментации.

ЭФП признакового уровня – это пятерка:

$$D = < w, \mu_D, s_D, y_D, A_D >, \tag{9}$$

где w – вес текущего речевого сегмента (2); μ_D – степень соответствия текущего речевого сегмента соответствующему признаку (значение функции принадлежности); s_D – состояние признакового ЭФП; y_D – выход признакового ЭФП; A_D – алгоритм, вычисляющий значение выхода признакового ЭФП:

$$A_D : w, \mu_D, s_D \rightarrow y_D \tag{10}$$

Алгоритм A_D включает три шага:

1. Накопление состояния. Состояние s_D ЭФП изменяется по формуле:

$$s_D(t) = s_D(t-1) \oplus \mu_D, \tag{11}$$

где запись $s_D(t)$ означает состояние ЭФП в момент времени t. Операция $\oplus$ – это операция накопления, формула которой заимствована из когнитивной модели логогена Мортона [13]:

$$a \oplus b = a + b - a \cdot b \tag{12}$$

2. Расчет выхода. Выход ЭФП вычисляется по формуле:

$$y_D = s_D \cdot w \tag{13}$$

3. «Забывание» некоторой порции информации. В качестве операции «забывания» выбрана операция вычитания:

$$s_D = \begin{cases} s_D - k_D & , \quad s_D > k_D \\ 0 & , \quad s_D \le k_D \end{cases} \tag{14}$$

В формуле (14) k_D – это коэффициент забывания (вещественное число из диапазона [0,1]). Все признаковые ЭФП должны регулярно уменьшать свои состояния на некоторые небольшие значения во избежание насыщения нескольких ЭФП к моменту принятия фонемного решения. В работе применяется значение $k_D = 0.15$.

Полезным в модели FCAS является учет фонетической статистики словаря (условных вероятностей появления фонологических признаков). В момент принятия фонемного

решения все признаковые ЭФП, которые будут работать на следующем этапе (речевом сегменте), получают дополнительную активацию, в зависимости от текущего звука речи. Эта дополнительная активация пропорциональна условным вероятностям появления конкретного признака после текущего признака:

$$s_{D_i} = s_{D_i} \oplus k_{PD} \cdot P(D_i \setminus D_{cur}), \tag{15}$$

где $P(D_i \setminus D_{cur})$ – условная вероятность появления признака D_i после текущего признака D_{cur}; k_{PD} – коэффициент пропорциональности.

ЭФП словесного уровня определяется как четверка:

$$W = < \{w_t\}_{t=1}^{n}, \{d_t\}_{t=1}^{n}, y_W, A_W >, \tag{16}$$

где $\{w_t\}_{t=1}^{n}$ – последовательность весов речевых сегментов в интервале между двумя моментами принятия фонемного решения на словесном уровне (иначе говоря, последовательность весов предполагаемых фонем в слове); $\{d_t\}_{t=1}^{n}$ – последовательность выходных значений признаковых ЭФП, соответствующих фонетическим признакам, присутствующим в данном слове; n – число фонем в слове, представленном данным ЭФП словесного уровня; y_W – выход словесного ЭФП; A_W – алгоритм, вычисляющий значение выхода словесного ЭФП:

$$A_W : \{w_t\}_{t=1}^{n}, \{d_t\}_{t=1}^{n} \to y_W. \tag{17}$$

Так как фонемные ЭФП в предлагаемой модификации модели FCAS отсутствуют, то алгоритм A_W состоит всего из одного шага – вычисления суммарного состояния ЭФП по признакам всех звуков в данном слове:

$$y_W = s_{WD} = \frac{1}{n} \sum_{t=1}^{n} w_t \cdot d_t. \tag{18}$$

Моментами принятия фонемного решения на уровне словесных ЭФП являются речевые сегменты пауз. Если максимальный выход среди всех ЭФП словесного уровня превышает порог k_W=0.5, то к выходной последовательности команд FCAS добавляется последовательность символов, представляющая команду, генерируемую данным ЭФП. В противном случае (команда отсутствует в лексиконе FCAS) система не выдает на выходе никакой команды.

Результаты экспериментов. Тестирование проводилось на роботах LEGO MindStorms NXT 2.0. Комплекс распознавания команд был развернут на отдельном компьютере, соединение которого с роботами производилось по Bluetooth (рис. 4).

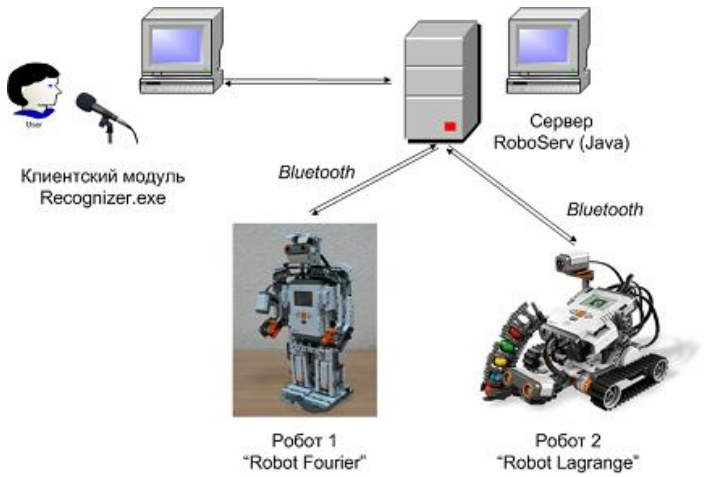

Рисунок 4 – Схема аппаратно-программного комплекса

Тестирование проводилось в стандартной компьютерной аудитории без звукоизоляции. Использовался бюджетный микрофон низкого качества, по аналогии с микрофонами, традиционно устанавливаемыми на роботах. Лексикон голосовых команд включал следующие словосочетания: «GO STRAIGHT», «GO BACK», «LEFT», «RIGHT», «STOP», «SPEED UP», «SLOW DOWN», «DANGER», «SEARCH», «TURN AROUND», «ROBOT FOURIER», «ROBOT LAGRANGE». В тестировании принимали участие 10 дикторов с голосами в диапазоне частот основного тона от 120 Гц до 200 Гц. Каждый диктор произнес все команды из лексикона по 50 раз. Результаты распознавания команд приведены в таблице 3, в которой в качестве D обозначено число удалений команд (число случаев, когда система не выдала никакой команды на выходе), S – число неверно распознанных команд, CER (Command Error Rate) – процент ошибок распознавания команды.

Таблица 3 – Результаты экспериментов по распознаванию голосовых команд

Команда	D	S	CER	Команда	D	S	CER
GO STRAIGHT	1	20	4.2	SLOW DOWN	1	24	5
GO BACK	2	36	7.6	DANGER	2	26	5.6
LEFT	5	41	9.2	SEARCH	0	19	3.8
RIGHT	7	37	8.8	TURN AROUND	1	25	5.2
STOP	4	28	6.4	ROBOT FOURIER	0	29	5.8
SPEED UP	2	25	5.4	ROBOT LAGRANGE	0	26	5.2

Как видно из табл. 3, показатель CER различается для разных слов, не превышая при этом 9.2% (CER команды «LEFT»); среднее значение CER составило 6.02%. Несмотря на ограниченность размера словаря и числа дикторов, данный показатель можно считать хорошим результатом, т.к. системе не требуется ни обучаться на многочасовых коллекциях речевых данных, ни подстраиваться под диктора. Кроме того, предложенная в статье конфигурация модели FCAS – лишь одна из многих возможных (блоки на рис. 1 являются взаимозаменяемыми).

Отдельно приведем также результаты автоматического взвешивания участков голосовых команд: средние значения степени ударности (3), разгармонизации (1) и веса (2) речевого сегмента (табл. 4).

Таблица 4 – Результаты экспериментов по взвешиванию речевых сегментов

Тип звуков речи	AM	INH	w
Невокализованные	0.16	0.87	0.87
Вокализованные ударные	0.82	0.08	0.82
Вокализованные безударные гласные	0.44	0.15	0.44
Вокализованные согласные	0.35	0.51	0.51

Как видно из табл. 4, параметры (1), (2) и (3) позволяют адекватно оценить важность отдельных типов звуков речи для обработки. Наименьший вес вполне ожидаемо имеют безударные гласные, т.к. им свойственна непостоянная спектральная картина, и их восприятие определяется, в основном, соседними звуками. Наибольший вес имеют невокализованные звуки, особенно шипящие, т.к. они однозначно воспринимаются человеком и характеризуются большими значениями *INH*. Наибольший вес среди вокализованных звуков имеют ударные гласные. Как видно из табл. 4, значение *AM*=0.82 можно использовать в качестве ориентира при автоматическом выделении ударений в словах.

Выводы

Описанная в статье легковесная модификация модели FCAS хорошо зарекомендовала себя при распознавании ограниченного набора голосовых команд. Благодаря нечеткой модели фонологической классификации звуков речи, системе удается более точно определять звуковой состав команд, а применение методики взвешивания речевых сегментов на основе просодических характеристик позволяет подчеркнуть наиболее важные для восприятия участки речи, уменьшить влияние трудно различимых участков в модели постобработки речевого сигнала и, следовательно, повысить эффективность распознавания команд в целом.

Модель FCAS можно рассматривать как перспективный вариант для построения робототехнических систем распознавания голосовых команд, а также для интеграции с современными статистическими моделями постобработки речи. Возможными направлениями усовершенствования модели является адаптивная настройка коэффициентов модели и робастные методы извлечения признаков.

Список литературы

1. Niculescu A. Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy / A.Niculescu, B. van Dijk, A.Nijholt, H.Li, S.L.See // International Journal of Social Robotics. – 2013. – Vol.5(2). – P.171-191.
2. Valin J. Robust Recognition of Simultaneous Speech by a Mobile Robot / J.Valin, S.Yamamoto, J.Rouat, F.Michaud, K.Nakadai, H.Okuno // IEEE Transactions on Robotics. – 2007. – Vol.23(4). – P.742-752.
3. Google Speech API Community Group [Электронный ресурс]. – Режим доступа: http://www.w3.org/community/speech-api/.
4. Apple Siri [Электронный ресурс]. – Режим доступа: http://www.apple.com/iphone/features/siri.html.
5. Huang X. Spoken Language Processing: A guide to theory, algorithm, and system development / X. Huang, A. Acero, H. Hon. – Prentice Hall. – 2001. – 980 p.
6. Controling a robot using voice – Speech recognition module for robots [Электронный ресурс]. – Режим доступа : http://www.generationrobots.com/en/content/59-speech-recognition-system-robot-parallax.
7. How to Build a Robot Tutorials – Society of Robots [Электронный ресурс]. – Режим доступа : http://www.societyofrobots.com/sensors_voice_recognition_robot.shtml.
8. Шарий Т.В. Модель постобработки речевых сигналов FCAS / Шарий Т.В. // Вісник Чернігівського державного технологічного університету. Серія «Технічні науки». – 2012. – № 4 (61). – С. 157-165.
9. Каргин А.А. Анализ речевых сигналов с учетом просодических характеристик / Каргин А.А., Шарий Т.В. // Сборник трудов X международной конференции «Интеллектуальный анализ информации ИАИ-2010», Киев. – 2010. – С.339-344.
10. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project / G. Peeters // CUIDADO Proj. Report. – 004. – 72 p.
11. Каргин А.А. Применение нечеткой логики в системах фонологической классификации звуков речи / Каргин А.А., Шарий Т.В. // Искусственный интеллект. – 2010. – № 3. – С. 210-219.
12. McClelland J.L. The TRACE Model of Speech Perception / J.L. McClelland, J.L. Elman // Cognitive Psychology. – Vol. 18. – 1986. – P.1-86.
13. Morton J. Word recognition / J. Morton, J.C. Marshall // Psycholinguistics 2: Structures and Processes. – 1979. – P.107-156.

References

1. Niculescu A. Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy / A.Niculescu, B. van Dijk, A. Nijholt, H. Li, S.L.See // International Journal of Social Robotics. – 2013. – Vol.5(2). – P.171-191.
2. Valin J. Robust Recognition of Simultaneous Speech by a Mobile Robot / J.Valin, S.Yamamoto, J.Rouat, F.Michaud, K.Nakadai, H.Okuno // IEEE Transactions on Robotics. – 2007. – Vol.23(4). – P.742-752.
3. Google Speech API Community Group [Electronic resource]. – Access mode: http://www.w3.org/community/speech-api/.

4. Apple Siri [Electronic resource]. – Access mode: http://www.apple.com/iphone/features/siri.html.
5. Huang X. Spoken Language Processing: A guide to theory, algorithm, and system development / X.Huang, A.Acero, H.Hon. – Prentice Hall. – 2001. – 980p.
6. Controling a robot using voice – Speech recognition module for robots [Electronic resource]. – Access mode: http://www.generationrobots.com/en/content/59-speech-recognition-system-robot-parallax.
7. How to Build a Robot Tutorials – Society of Robots [Electronic resource]. – Access mode: http://www.societyofrobots.com/sensors_voice_recognition_robot.shtml.
8. Sharii T.V. The FCAS Model of Speech Signal Post-Processing / T.V. Sharii // Visnyk Chernihivs'kogo derjavnogo technologitchnogo universytetu. Seriya «Technitchni nauky». – 2012. – №4(61). – P.157-165.
9. Kargin A.A. An Analysis of Speech Signals Based on Prosodic Features / A.A. Kargin, T.V. Sharii // Proceedings of X International Conference «Intellectual Analysis of Information IAI-2010», Kyiv. – 2010. – P.339-344.
10. Peeters G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project / G. Peeters // CUIDADO Proj. Report. – 2004. – 72 p.
11. Kargin A.A. The Use of Fuzzy Logic in the Systems of Phonological Classification of Speech Sounds / Kargin A.A., Sharii T.V. // Iskusstvenniy Intellekt. – 2010. – №3. – P.210-219.
12. McClelland J.L. The TRACE Model of Speech Perception / J.L.McClelland, J.L.Elman // Cognitive Psychology. – Vol.18. – 1986. – P.1-86.
13. Morton J. Word recognition / J.Morton, J.C.Marshall // Psycholinguistics 2: Structures and Processes. – 1979. – P.107-156.

RESUME

T.V. Sharii

A Mobile Robot Voice Control Based on the Fcas Cognitive Model

The article considers issues relating to use of the FCAS fuzzy cognitive model [8] for a mobile robot voice control commands recognition. The lightweight modification of the model is developed that allows to operate a robot effectively in real-time mode.

The key element of proposed model is the FCAS kernel that processes such heterogeneous information as the relative weight of a speech sound, the membership of a speech sound to various phonological classes and the phonetic statistics of robot's lexicon.

The automatic speech segment weighting block is based on analysis of prosodic features of a speech signal allowing to emphasize the most important speech fragments for human perception and to reduce the influence of poorly distinguishable fragments on recognition accuracy. The differential feature block rests upon the fuzzy model of phonological classification that helps to determine a sound content of commands more accurately. The FCAS kernel is a set of feature-level and word-level elementary phonetic processors accumulating the heterogeneous information obtained at the speech parameterization stage.

The special software bundle has been built and tested on LEGO MindStorms NXT 2.0 robots with lexicon containing 12 commands. A total of 10 speakers have been involved in the testing. An average command recognition error rate is 6%.

Статья поступила в редакцию 03.04.2014.