УДК 004.934
ББК 30в6

О ЗАВИСИМОСТИ ТОЧНОСТИ МЕТОДА НЕЧЕТКОГО ФОНЕТИЧЕСКОГО КОДИРОВАНИЯ-ДЕКОДИРОВАНИЯ СЛОВ ОТ ДЛИТЕЛЬНОСТИ КОМАНДЫ В СИСТЕМАХ ГОЛОСОВОГО УПРАВЛЕНИЯ

Савченко Л. В.[1]
(*Нижегородский Государственный Лингвистический Университет, Нижний Новгород*)

Рассматривается задача распознавания изолированных слов русского языка для систем голосового управления на основе метода нечеткого фонетического кодирования-декодирования. Предложена модификация выражения для определения степени принадлежности звука к эталонным фонемам. Представлены экспериментальные исследования точности метода в задаче распознавания голосовых команд в зависимости от длины команды. Показано, что предложенный подход характеризуется более низкой вероятностью ошибки распознавания по сравнению с оригинальным методом.

Ключевые слова: автоматическое распознавание речи, принцип минимума информационного рассогласования, теория нечетких множеств, метод нечеткого фонетического кодирования-декодирования слов.

1. Введение

Направление автоматического распознавания речи (АРР) и, в частности, распознавание изолированных слов в задаче голосового управления, в настоящее время становится все более актуальным [3]. К сожалению, существующие коммерческие

[1] *Савченко Людмила Васильевна, аспирант* (*lyudmilasavchenko @yandex.ru*).

решения, особенно для русского языка, имеют недостаточную точность распознавания для их практического применения в прикладных областях с повышенными требованиями к надежности: при наличии акустических помех, изменении голоса, стиля произношения, физического состояния говорящего, аддитивного фонового шума [2, 4]. Для повышения точности распознавания может использоваться метод фонетического декодирования слов (ФДС) [5], основанный на принципе минимума информационного рассогласования Кульбака-Лейблера [8]. В нем задача распознавания слов сводится к выделению границ слогов и распознаванию гласной фонемы в каждом слоге. Основными преимуществами метода ФДС над другими методами распознавания являются быстрая настройка на голос конкретного диктора и автоматически перенастраиваемый рабочий словарь. К сожалению, точность метода ФДС иногда оказывается недостаточной, так как близкие по звучанию фонемы часто объединяют в один кластер и на выходе алгоритма распознавания присутствует много альтернативных решений [4]. Ранее [6] на основе метода ФДС и теории нечетких множеств для задачи распознавания слов был предложен метод нечеткого фонетического кодирования-декодирования слов (НФКДС). В настоящей работе исследуется точность распознавания метода в зависимости от длительности команды**.** Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области АРР.

2. *Метод нечеткого фонетического кодирования-декодирования слов*

Пусть задано множество из L>1 эталонных слов $\{X_l\}$, где $l=\overline{1,L}$ – номер слова-эталона. Согласно фонетическому подходу [1], каждое эталонное слово разбивается на последовательность фонем (транскрипцию) $X_l=\{r_{l,1},r_{l,2},...,r_{l,L_l}\}$. Здесь L_l – длительность слова/словосочетания (в фонемах), а числа $r_{l,j}\in\{1,...,R\}$ – номера фонем из фонетического алфавита

$\left\{\mathbf{x}_r^*\right\} r=\overline{1,R}$, где R - количество фонем в алфавите; $\mathbf{x}_r^*$ – вектор отсчетов сигнала r-го эталонного звука. Задача состоит в том, чтобы поступившей на вход голосовой команде X поставить в соответствие наиболее близкое к нему слово-эталон X_l.

Для задачи распознавания слов (голосовых команд) предполагаем, что входное слово X разбито на N слогов, причем границы каждого n-го слога ($n=\overline{1,N}$) определены с точностью до номера квазистационарного фрейма $\left(t_n^{(1)}, t_n^{(2)}\right)$. Далее для каждого n-го слога производится распознавание только среди гласных звуков, т.е. фонетический алфавит $\left\{\mathbf{x}_r^*\right\} r=\overline{1,R}$ состоит из эталонов гласных фонем. Каждому слогу следует поставить в соответствие одну из R эталонных минимальных звуковых единиц (МЗЕ). Вначале слог разбивается на непересекающиеся фреймы длиной $\tau\approx 0{,}01\ldots 0{,}03$ с, где T - общее число фреймов в анализируемом речевом сигнале. После этого каждый полученный парциальный сигнал $\mathbf{x}(t)=\|x_1(t)\ldots x_M(t)\|$ (здесь $M=\tau F$ - количество отсчетов во фрейме, F – частота дискретизации сигнала) рассматривается в пределах конечного списка эталонных МЗЕ и отождествляется с той $\mathbf{x}_{\nu(t)}$ из них, которая отвечает принципу минимума рассогласования Кульбака-Лейблера между сигналом $\mathbf{x}(t)$ и эталоном $\mathbf{x}_r^*$:

$$(1)\quad \nu(t)=\arg\min_{r=\overline{1,R}} \rho_{KL}\left(\mathbf{x}(t),\mathbf{x}_r^*\right),\ t=\overline{1,T},$$

где

$$(2)\quad \rho_{KL}(\mathbf{x}(t),\mathbf{x}_r^*)=\frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_r(f)}-\ln\frac{G_x(f)}{G_r(f)}\right)-1.$$

Это известная формулировка критерия минимума информационного рассогласования на основе авторегрессионной (АР) модели речевого сигнала [9]. Здесь $G_x(f)$– выборочная оценка спектральной плотности мощности (СПМ) входного сигнала как

функция дискретной частоты f , а $G_r(f)$ – СПМ эталона r-ой фонемы, F – верхняя граница частотного диапазона речевого сигнала или используемого канала связи.

Согласно методу ФДС, каждой МЗЕ $\mathbf{x}_r^*$ ставится в соответствие некий числовой код $c(r)\in\{1, ..., C\}$, где в общем случае $C\le R$. Для каждого фрейма в момент времени t решение принимается по принципу минимума информационного рассогласования. Итоговое решение принимается в пользу наиболее часто встречающегося кода c*:

$$(3)\quad c^* = \arg\max_{c=\overline{1,C}} \sum_{t=1}^{T} \delta\big(c(\nu(t)) - c\big),$$

где $\delta(x)$ - дискретная дельта-функция, а $\nu(t)$ определяется согласно (1).

Недостаток метода ФДС, упомянутый во введении, может быть устранен с помощью предложенного нами ранее метода НФКДС [6], в котором каждому слогу ставится в соответствие нечеткое множество $\left\{\left(\mathbf{x}_r^*, \mu_j\left(\mathbf{x}_r^*\right)\right)\right\}$, где $\mu_j\left(\mathbf{x}_r^*\right)$ - степень принадлежности эталона $\mathbf{x}_r^*$ к j-й МЗЕ, определяемая как $\mu_j\left(\mathbf{x}_r^*\right) = P\left(\mathbf{x}_j^* / \mathbf{x}_r^*\right)$. Для оценки условной вероятности принадлежности к j-й фонеме воспользуемся известным свойством [8]: рассогласование Кульбака-Лейблера между объектами классов j и r в асимптотике с точностью до постоянного множителя λ=const>0 имеет нецентральное хи-квадрат распределение с числом степеней свободы, определяемым количеством независимых параметров классифицируемого объекта ($K=M-p$ для АРР и гауссова сигнала) и параметром нецентральности $\lambda \cdot \rho\left(\mathbf{x}_j^* / \mathbf{x}_r^*\right)$. Тогда, если M достаточно велико, воспользуемся центральной предельной теоремой и определим вероятность из

известного распределения минимума независимых нормальных величин [7] следующим образом:

$$\mu_j\left(\mathbf{x}_r^*\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\left(-\frac{t^2}{2}\right)\prod_{\substack{i=1\\ i\neq j}}^{R}\left(\frac{1}{2}-\right.$$

$$\left.-\Phi\left(\frac{t\cdot\sqrt{8\lambda\cdot\rho(\mathbf{x}_j^*,\mathbf{x}_r^*)+p-1}+2\lambda\cdot\left(\rho(\mathbf{x}_j^*,\mathbf{x}_r^*)-\rho(\mathbf{x}_j^*,\mathbf{x}_r^*)\right)}{\sqrt{8\lambda\cdot\rho(\mathbf{x}_j^*,\mathbf{x}_r^*)+p-1}}\right)\right)dt. \quad (4)$$

Каждому фрейму входного сигнала $\mathbf{x}(t)$ также ставится в соответствие нечеткое множество вида $\left\{\left(\mathbf{x}_r^*,\mu\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)\right)\right\}$. Мы предполагаем, что степень принадлежности $\mu\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)$ определяется как апостериорная вероятность принадлежности фрейма $\mathbf{x}(t)$ к r-й гласной [8]:

$$\mu\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)=\frac{\exp\left(-\lambda\cdot\rho\left(\mathbf{x}(t),\mathbf{x}_r^*\right)\right)}{\sum_{i=1}^{R}\exp\left(-\lambda\cdot\rho\left(\mathbf{x}(t),\mathbf{x}_i^*\right)\right)}, \quad (5)$$

где $\rho(\cdot)$ – произвольное рассогласование, λ – параметр масштабирование, который подбирается экспериментально для каждой конкретной меры близости.

Далее, используя операцию нечеткого пересечения множеств ближайшей эталонной фонемы и фрейма входного сигнала, получим результирующее множество $\left\{\mathbf{x}_r^*,\mu(r,t)\right\}$:

$$\mu(r,t)=\min\left(\mu_{\nu(t)}\left(\mathbf{x}_r^*\right),\mu\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)\right), \quad (6)$$

где $\nu(t)$ определяется согласно (1). На основе всех $\mu(r, t)$ каждому слогу ставится в соответствие нечеткое множество

$\left\{\left(\mathbf{x}_r^*, \mu(r)\right)\right\}$, где применяется простое голосование: $\mu(r) = \frac{1}{T}\sum_{t=1}^{T}\mu(r,t)$.

Формула (4) основана на том, что рассогласование Кульбака-Лейблера между объектами разных классов в асимптотике имеет нецентральное хи-квадрат распределение. К сожалению, такое предположение в практических приложениях не всегда справедливо в связи с известной вариативностью устной речи. Поэтому в настоящей работе по аналогии с (5) предлагается упрощение формулы (4):

$$(7)\quad \mu_j\left(\mathbf{x}_r^*\right) = \frac{\exp\left(-\lambda\cdot\rho\left(\mathbf{x}_j^*, \mathbf{x}_r^*\right)\right)}{\sum_{i=1}^{R}\exp\left(-\lambda\cdot\rho\left(\mathbf{x}_j^*, \mathbf{x}_i^*\right)\right)}.$$

Такое упрощение хорошо согласуется с предположением о том [8], что, если $\mathbf{x}(t)$ принадлежит тому же классу, что и эталон $\mathbf{x}_\gamma^*$, то в асимптотике

$$(8)\quad \forall r \in \{1,\ldots,R\} \qquad \rho\left(\mathbf{x}(t), \mathbf{x}_r^*\right) \to \rho\left(\mathbf{x}_\gamma^*, \mathbf{x}_r^*\right).$$

Поэтому, если $\gamma=\nu(t)$, то $\mu(\nu(t), t)\approx 1$. В противном случае, если $\gamma\neq\nu(t)$, то $\mu(\nu(t),t) = \mu(\mathbf{x}(t)/\mathbf{x}_{\nu(t)}^*) << 1$. Таким образом, выражение (6) приводит к существенному понижению степеней принадлежности в случае ошибки распознавания ($\gamma\neq\nu(t)$).

Возвращаясь к исходной задаче распознавания слов n-му слогу ставится в соответствие последовательность частот $\mu_n(r), r=\overline{1,R}$, где

$$(9)\quad \mu_n(r) = \frac{1}{t_n^{(2)} - t_n^{(1)} + 1}\cdot\sum_{t=t_n^{(1)}}^{t_n^{(2)}}\mu(r,t).$$

Далее для каждого слова-эталона X_l оценивается ее корреляция с распознаваемым речевым сигналом

$$(10)\ \mu_l = \begin{cases} \sum_{n=1}^{N} \mu_n\left(r_{l,n}\right), & L_l = N \\ 0 & L_l \neq N \end{cases}.$$

Тогда решение задачи АРР принимается в пользу слова X^* по следующему критерию [11]

$$(11)\ X^* = \underset{X_l, l=\overline{1,L}}{\arg\max}\ \mu_l .$$

Система выражений (5)-(11) и определяет алгоритм распознавания изолированных слов на основе модифицированного метода НФКДС.

3. Результаты экспериментальных исследований в задаче распознавания слов

Проведем сравнение точности распознавания методов ФДС, НФКДС, а также известных систем распознавания Google Voice Search [10] и CMU Sphinx [12] для словаря лекарств (1913 наименований) и продуктов питания (300 наименований). Для тестирования диктор произносил 300 слов/словосочетаний из словаря с четким выделением слогов (каждое слово произносилось по 3 раза). В качестве эталонов брались десять гласных звуков этого же диктора по три реализациям (для каждой МЗЕ). Методы ФДС и НФКДС использовались совместно с рассогласованием Кульбака-Лейблера (2), порядок АР-модели p=15 и Евклида с MFCC коэффициентами, отношение сигнал/шум 20 дБ, для рассогласования Кульбака-Лейблера оптимальный параметр масштабирования λ=0,05, а для MFCC признаков λ=1. Оценка вероятности ошибки распознавания (в %) в зависимости от количества слогов для различных мер близости и словарей представлена на рис.1,2.

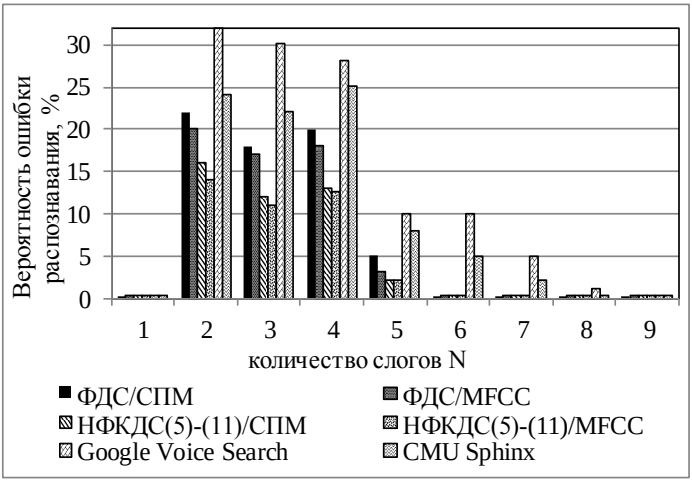

Рис. 1. Зависимость вероятности ошибки распознавания слов от количества слогов N для словаря «Лекарства»

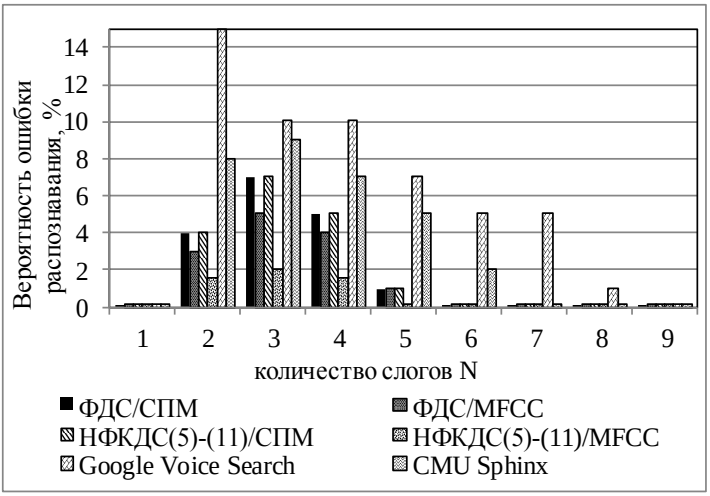

Рис. 2. Зависимость вероятности ошибки распознавания слов от количества слогов N для словаря «Продукты»

Как видно из этих рисунков метод НФКДС превосходит по точности распознавания метод ФДС. Так, например, вероятность ошибки распознавания метода НФКДС для слов, состоящих из трех слогов для рассогласования Кульбака-Лейблера между оценками СПМ сигналов, составляет 12%, что на 6% ниже аналогичного показателя для метода ФДС, на 18% и 10% ниже аналогичного показателя для известной системы APP Google Voice Search и CMU Sphinx [11] соответственно (рис. 1). Качество известной системы Google Voice Search оказалось невысоким для словаря «Лекарства», так как названия многих лекарств является специфическим и отсутствует в универсальном словаре системы, в то же время точность системы CMU Sphinx оказалось высоким, т.к. система использует возможность загрузки словаря. Для оригинального метода НФКДС ошибка распознавания на 1,5-3% выше, чем для предложенного НФКДС (5)-(11), таким образом, упрощение метода по формуле (7) позволяет повысить точность распознавания.

4. Заключение

В работе проведено сравнение точности распознавания модифицированного метода НФКДС (5)-(11), основанного на методе ФДС и теории нечетких множеств, в зависимости от длины распознаваемого слова (голосовой команды) с традиционными методами и системами APP, такими как метод ФДС, системой распознавания Google Voice Search и CMU Sphinx. Экспериментально продемонстрировано, что точность распознавания длинных слов превосходит точность распознавания коротких (рис. 1, 2), что связано с тем, что на выходе алгоритма распознавания для длинных слов практически отсутствуют слова с одинаковой степенью принадлежности. Методы ФДС и НФКДС превосходят по точности распознавания известные системы APP за счет послогового произношения и настройки на пользователя (дикторозависимый режим). Точность распознавания предложенного метода НФКДС совместно с MFCC признаками на 0,5-2% выше, чем для рассогласования Кульбака-Лейблера.

Литература

1. НИЦЕНКО А.В. *Алгоритмы пофонемного распознавания слов наперед заданного словаря* / Ниценко А. В., Шелепов В. Ю. // Искусственный интеллект. – 2004. – № 4. – С. 633–639.
2. ПОТАПОВА Р.К. *Перспективы развития прикладного речеведения* // Речевые технологии. -2008. -№1. – С. 5-17.
3. РОНЖИН А. Л. *Метод автоматического распознавания голосовых команд и неречевых акустических событий* /Ронжин А. Л., Глазков С. В. // Информационно-управляющие системы. – 2012. – № 4. – С. 74 –77.
4. САВЧЕНКО А.В. *Адаптивный алгоритм распознавания речи на основе метода фонетического декодирования слов в задаче голосового управления* // Информационные технологии. – 2013. – №4. – С. 34-39.
5. САВЧЕНКО В.В. *Метод фонетического декодирования слов в задаче автоматического распознавания речи на основе принципа минимума информационного рассогласования* // Известия вузов России. Радиоэлектроника. – 2009. – Вып.5. – С. 31-41.
6. САВЧЕНКО Л.В. *Алгоритм пофонемного распознавания устной речи на основе метода нечеткого фонетического кодирования-декодирования слов* // Информационно-управляющие системы. – 2014. – №1. – С. 23-31.
7. HILL, J. E. The Minimum of n Independent Normal Distributions. http://www.untruth.org/~josh/math/normalmin.pdf (дата обращения: 03.03.2014).
8. KULLBACK, S. Information Theory and statistics / S. Kullback / /Dover Pub. – 1997. – 399 p.
9. MARPLE, S. L. (Jr.). Digital spectral analysis and with application / S. L. Marple / /Englewood Cliffs. – New Jersey: Prentice-Hall. – 1987. – 492 p.
10. SCHUSTER, M. Speech recognition for mobile devices at Google / M. Schuster // Proceedings of the 11th Pacific Rim international conference on Trends in Artificial Intelligence, LNCS.–2010.– Vol. 6230.– P. 8–10.

11. ZADEH, L. A. Fuzzy Sets / L. F. Zadeh / / Information Control. – 1965. – Vol. 8. – P. 338–353.
12. http://cmusphinx.sourceforge.net/ (дата обращения: 28.10.2013).

DEPENDENCE OF THE ACCURACY OF FUZZY PHONETIC CODING-DECODING METHOD ON THE COMAND DURATION IN VOICE CONTROL APPLICATIONS

Lyudmila Savchenko, Nizhniy Novgorod State Linguistic University, Nizhniy Novgorod, graduate student (lyudmilasavchenko@mail.ru).

Abstract: *The problem of isolated words recognition for Russian language in voice control applications on the basis of the fuzzy phonetic coding-decoding method is explored. The modification of the definition the grade of sound membership to model phoneme is proposed. The experimental results of the dependence of the method's accuracy on the command length are presented. It is shown that the proposed approach is characterized by lower error rate in comparison with the original method.*

Keywords: automatic speech recognition, minimum information discrimination principle, fuzzy set theory, fuzzy phonetic coding-decoding method.