УДК 004.934

АЛГОРИТМ ПОФОНЕМНОГО РАСПОЗНАВАНИЯ УСТНОЙ РЕЧИ НА ОСНОВЕ МЕТОДА НЕЧЕТКОГО ФОНЕТИЧЕСКОГО КОДИРОВАНИЯ-ДЕКОДИРОВАНИЯ СЛОВ

Л. В. Савченко[a, 1], *аспирант*
[a]*Нижегородский государственный лингвистический университет им. Н. А. Добролюбова, Нижний Новгород, РФ*

Цель исследования: *повышение точности автоматического распознавания русской речи в системах голосового управления.* ***Методы:*** *предложена модификация метода нечеткого фонетического кодирования-декодирования слов, использующая известные признаки согласных звуков, которые классифицируются с помощью алгоритмов машинного обучения на основе приближенных множеств и деревьев решений. Приведены наиболее характерные правила классификации (ЕСЛИ..., ТО ...) для каждого типа звука.* ***Результаты:*** *представлены результаты экспериментального исследования в задаче распознавания голосовых команд для широко используемых в автоматической обработке речи мер близости (Кульбака — Лейблера, Евклида, Spectral distortion) совместно с популярными признаками речевого сигнала (оценки спектральных плотностей мощности, коэффициенты линейного предсказания, кепстральные коэффициенты). Показано, что точность распознавания речи для предложенной модификации на 3—7, 1—7 и 1,5—24 % выше точности известного аналога — метода фонетического декодирования слов, современной библиотеки CMU Sphinx и популярной системы Google Voice Search соответственно.* ***Практическая значимость:*** *увеличение степени принадлежности входного сигнала к эталонному слову за счет предложенного алгоритма классификации согласных звуков приводит к увеличению точности и к уменьшению количества альтернативных решений на выходе алгоритма распознавания.*

Ключевые слова *— автоматическое распознавание изолированных слов, нечеткие множества, приближенные множества, деревья решений, метод нечеткого фонетического кодирования-декодирования слов.*

Введение

Задача распознавания изолированных слов/словосочетаний в области автоматического распознавания речи (АРР), в частности для систем голосового управления, в настоящее время становится все более актуальной [1—3]. Эффективным инструментом построения голосового интерфейса с быстрой настройкой на нового диктора и автоматической адаптацией словаря [2] является метод фонетического декодирования слов (ФДС) [4], основанный на пофонемном распознавании изолированно произнесенных фраз [5]. К сожалению, метод ФДС характеризуется невысокой точностью распознавания для коротких по длительности слов [4], что связано в первую очередь с тем, что близкие по звучанию звуки объединяются в один кластер. Для повышения точности распознавания нами был предложен метод нечеткого фонетического кодирования-декодирования слов (НФКДС) [6], в котором каждой минимальной звуковой единице (МЗЕ) ставится в соответствие не один информационный центр-эталон, как в методе ФДС, а нечеткое множество эталонных МЗЕ. К сожалению, методы ФДС и НФКДС для распознавания используют только гласные звуки (так как точность их распознавания наиболее высока [3]). В настоящей работе для повышения точности распознавания предложена модификация НФКДС, использующая признаки согласных звуков для уточнения решения. Полученные результаты и сделанные по ним выводы рассчитаны на широкий круг специалистов в области обработки и распознавания речи.

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, каждое эталонное слово разбивается на последовательность фонем (транскрипцию) $X_l=\left\{r_{l,1}, r_{l,2}, \ldots, r_{l,L_l}\right\}$ [4]. 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 поступившему на вход речевому сигналу (слову) X поставить в соответствие наиболее близкое к нему слово-эталон X_l.

Сведем задачу распознавания изолированных слов/словосочетаний к распознаванию гласного звука в слоге — поступающему на вход речевому сигналу $\mathbf{x}$ с частотой дискретизации F [Гц] следует поставить в соответствие одну из R эталонных МЗЕ. Вначале $\mathbf{x}$ разбивается на непересека-

[1] Научный руководитель – кандидат технических наук, доцент, заместитель заведующего кафедрой математики и информатики Нижегородского лингвистического университета им. Н. А. Добролюбова *Д. Ю. Акатьев.*

ющиеся фреймы $\{\mathbf{x}(t)\}$, $t=\overline{1,T}$ длиной $\tau\approx 0{,}01...0{,}03$ с, где T — общее число фреймов в анализируемом речевом сигнале. После этого каждый полученный парциальный сигнал $\mathbf{x}(t)=\| x_1(t)... x_M(t)\|$ (здесь $M=\tau F$ — количество отсчетов в сегменте) рассматривается в пределах конечного списка МЗЕ $\left\{\mathbf{x}_r^*\right\}$ и отождествляется с той $\mathbf{x}_{\nu(t)}$ из них, которая отвечает принципу минимума некоторого рассогласования $\rho\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)$ между сигналом $\mathbf{x}(t)$ и эталоном $\mathbf{x}_r^*$:

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

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^*:

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

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

Для повышения точности распознавания обычно близкие между собой звуки (для русского языка, например, пары звуков, соответствующие произнесенным буквам {а, я}, {у, ю}, {о, ё}, {э, е}, {ы, и}) объединяют в один кластер. Такое объединение приводит к значительному сокращению количества различимых МЗЕ и, как следствие, к увеличению числа альтернативных решений на выходе алгоритма АРР, особенно для коротких по длительности слов [2].

Предположим, что в качестве меры близости применяется рассогласование Кульбака — Лейблера [7] для сопоставления отсчетов спектральных плотностей мощности (СПМ), оцененных на основе авторегрессионной (АР) модели речевого сигнала порядка p [8]. Тогда отмеченный недостаток 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 $P\left(\mathbf{x}_j^*/\mathbf{x}_r^*\right)$ принадлежности $\mathbf{x}_r^*$ к j-й фонеме воспользуемся известным свойством [7]: рассогласование Кульбака — Лейблера между объектами классов j и r в асимптотике с точностью до постоянного множителя $\alpha=\text{const}>0$ имеет нецентральное хи-квадрат распределение с числом степеней свободы, определяемым количеством независимых параметров классифицируемого объекта ($K=M-p$ для АРР и гауссова сигнала [4]) и параметром нецентральности $\alpha\cdot\rho\left(\mathbf{x}_j^*/\mathbf{x}_r^*\right)$. Тогда, если M достаточно велико, воспользуемся центральной предельной теоремой и определим вероятность $P\left(\mathbf{x}_j^*/\mathbf{x}_r^*\right)$ из известного распределения минимума независимых нормальных величин [9] следующим образом [10]:

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

$$\left.-\Phi\left(\frac{t\sqrt{8\alpha\cdot\rho\left(\mathbf{x}_r^*/\mathbf{x}_j^*\right)+p-1}+2\alpha\left(\rho\left(\mathbf{x}_r^*/\mathbf{x}_j^*\right)-\rho\left(\mathbf{x}_r^*/\mathbf{x}_i^*\right)\right)}{\sqrt{8\alpha\cdot\rho\left(\mathbf{x}_r^*/\mathbf{x}_i^*\right)+p-1}}\right)\right)\mathrm{d}t,$$

где $\Phi(\cdot)$ — интеграл вероятностей. 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 $P\left(\mathbf{x}_r^*/\mathbf{x}(t)\right)$ принадлежности фрейма $\mathbf{x}(t)$ к r-й гласной [7]:

$$\mu(\mathbf{x}(t)/\mathbf{x}_r^*)=P\left(\mathbf{x}_r^*/\mathbf{x}(t)\right)=\frac{\exp\left(-\alpha\cdot\rho\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)\right)}{\sum_{i=1}^{R}\exp\left(-\alpha\cdot\rho\left(\mathbf{x}(t)/\mathbf{x}_i^*\right)\right)}. \qquad (3)$$

Последнее выражение численно эквивалентно известной оценке апостериорной вероятности на выходе вероятностной нейронной сети [11].

Далее, используя операцию нечеткого пересечения множеств $\left\{\left(\mathbf{x}_r^*,\,\mu_j\left(\mathbf{x}_r^*\right)\right)\right\}$ и $\left\{\left(\mathbf{x}_r^*,\,\mu(\mathbf{x}(t)/\mathbf{x}_r^*\right)\right\}$, получаем результирующее множество $\left\{\mathbf{x}_r^*,\,\mu(r,t)\right\}$:

$$\mu(r,\,t)=\min\left(\mu_{\nu(t)}\left(\mathbf{x}_r^*\right),\,\mu\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)\right), \qquad (4)$$

где $\nu(t)$ определяется согласно (1). Операция нечеткого пересечения (4) приводит к существенному понижению степеней принадлежности фонем в случае ошибки распознавания, таким образом, их вклад в результирующее решение будет незначительным. 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: $\mu(r)=\frac{1}{T}\sum_{t=1}^{T}\mu(r,\,t)$. В свою очередь при распознавании слов получается нечеткое множество $\{(X_l,\,\mu_l)\}$, где степень принадлежности μ_l определяется как среднее арифметическое степеней принадлежности сло-

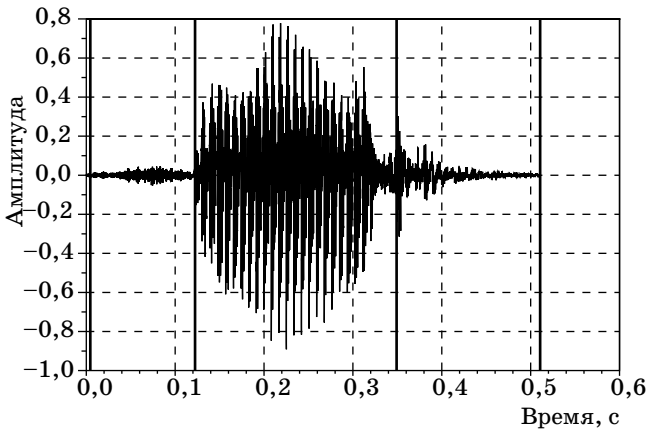
гов $\mu(r)$, составляющих слово. Итоговое решение принимается в пользу слова X^* с максимальной степенью принадлежности μ_l (так называемая дефаззификация [12]).

Автоматическое выделение согласных звуков

Как говорилось во введении, для метода ФДС (и, соответственно, НФКДС) фонетический алфавит $\{\mathbf{x}_r^*\}$ состоит только из гласных, а остальные фонемы игнорируются [13]. В настоящей работе для повышения точности распознавания слов предлагается выделять в анализируемом речевом сигнале не только гласные, но и согласные. Для этого распознаваемый речевой сигнал разбивается на стационарные сегменты (фонемная сегментация) [14, 15]. Предположим, что известен способ классификации таких сегментов на множестве классов (гласный, сонорный и т. п.). Если каждой фонеме слова-эталона X_l по его текстовому представлению поставить в соответствие определенный тип (широкий фонетический класс в терминологиях работы [16]), то после распознавания гласных в слогах входного слова на основе метода НФКДС (1)—(4) выделенная в распознаваемом слове последовательность типов сегментов сопоставляется с последовательностью типов фонем эталонной фразы X_l. Количество совпадений фонетических классов k_l умножается на некоторое фиксированное значение $\Delta\mu = \text{const}$, в результате слову/словосочетанию X_l ставится в соответствие нечеткое множество $\{(X_l, \tilde{\mu}_l)\}$, где

$$\tilde{\mu}_l = \begin{cases} \mu_l + k_l\Delta\mu, & \mu_l + k_l\Delta\mu < 1 \\ 1, & \mu_l + k_l\Delta\mu \geq 1 \end{cases}. \qquad (5)$$

Итоговое решение модифицированного таким образом метода НФКДС принимается в пользу слова X^* с максимальной степенью принадлежности $\tilde{\mu}_l$. Таким образом, ожидается, что такая классификация всех однородных сегментов входного слова должна привести к уменьшению числа слов с одинаковой степенью принадлежности и, в свою очередь, к повышению точности распознавания.



■ ***Рис. 1.*** Временная диаграмма речевого сигнала «шар»

Таким образом, задача сводится к классификации квазистационарных (однородных) сегментов, выделенных с помощью одного из известных алгоритмов фонемной сегментации [14]. Рассмотрим следующие типы (классы) фонем русского языка по участию в их образовании тона и шума [17]:

1) гласные, чьей акустической основой является только тон;

2) сонорные согласные (м, н, л, р), акустической основой которых является тон, шум практически отсутствует, по этому признаку они наиболее близки к гласным;

3) звонкие (б, в, д, г, з, й), в которых тон преобладает над шумом;

4) глухие (п, к, с, т, ц, ф, х), в которых шум преобладает над тоном;

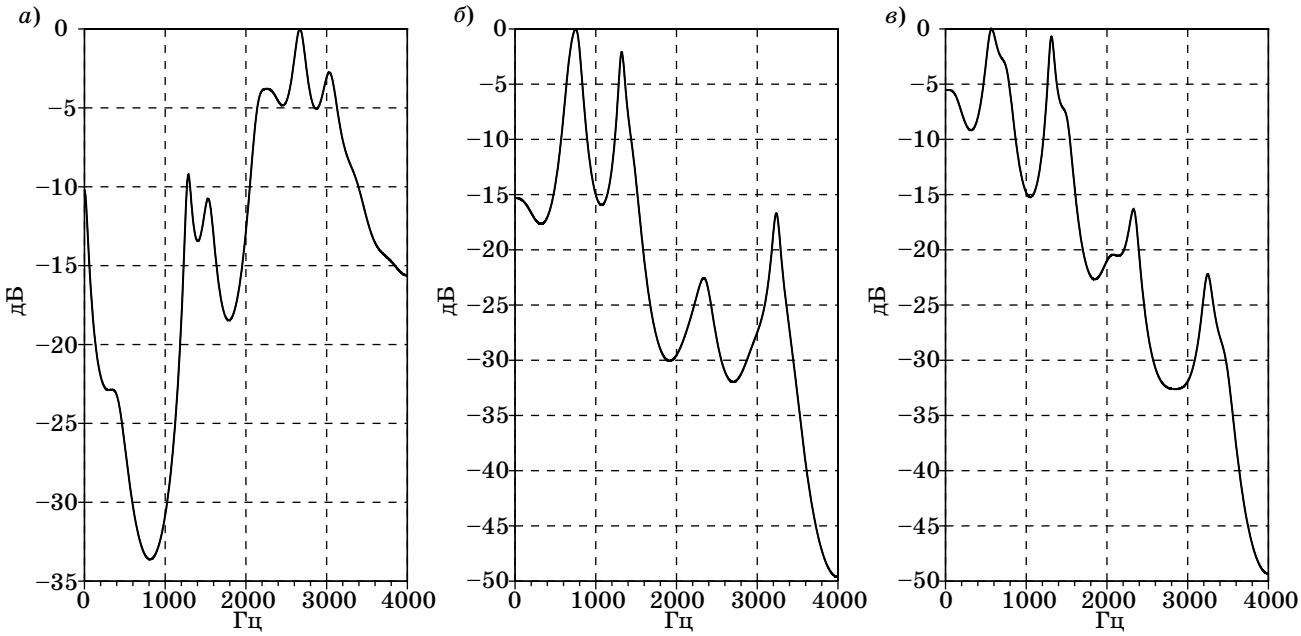
5) шипящие (ш, щ, ж, ч), чьей акустической основой является только шум.

Для всех пяти классов были выявлены три наиболее характерных признака: длительность $t_{\text{сегм}}$ [мс], отношение максимальной амплитуды фонемы к максимальной амплитуде содержащего ее слога $X_{\max}$ и частота $F_{\max}$ [Гц], при которой СПМ принимает максимальное значение [18].

В качестве примера на рис. 1 приведена временная диаграмма речевого сигнала для слова «шар», где черные вертикальные линии показывают границы звуковых сегментов «ш», «а» и «р», автоматически выделенных с помощью алгоритма сегментации [14], а на рис. 2 представлена зависимость авторегрессионной оценки СПМ от частоты для выделенных фонем «ш», «а» и «р».

По длительности $t_{\text{сегм}}$ гласная фонема «а» в 1,5 раза превосходит сонорную фонему «р» и более чем в 3 раза превосходит шипящую «ш» (см. рис. 1). Также можно заметить, что значение амплитуды $X_{\max}$ гласной фонемы принимает наибольшее значение, а шипящей — наименьшее. Из рис. 2 видно, что частота $F_{\max}$ для гласной и сонорной фонем принадлежит низкому частотному диапазону ($F_{\max} < 700$ Гц), в то время как для шипящей фонемы $F_{\max} = 2630$ Гц (верхний частотный диапазон).

Для нахождения правил вида ЕСЛИ..., ТО... классификации фонем по выбранным нами признакам применялись два алгоритма машинного



■ ***Рис. 2.*** Оценка СПМ фонемы «ш» (*а*), фонемы «а» (*б*) и фонемы «р» (*в*)

обучения: LERS (Learning from Examples Based on Rough Sets) [19] и CART (Classification and Regression Tree) [20]. Для этого шестью дикторами (тремя мужчинами и тремя женщинами) были записаны по 1000 реализаций изолированно произнесенных звуков русского языка (по 200 реализаций на каждый класс, включая гласные), из которых автоматически сформирован текстовый файл, содержащий значения трех признаков и истинный класс звука. Этот файл использовался для обучения алгоритмов классификаторов LERS (программа Rough Set Exploration System 2.2.2) и CART (Deductor Studio Academic 5.2). На выходе алгоритм LERS выдал набор из 17 правил, а алгоритм CART — из 18 правил, наиболее характерные из которых приведены в табл. 1, откуда видно, что применяемые нами признаки действительно приводят к следующей классификации звуков, а именно:

— гласные звуки характеризуются наибольшим значением $t_{\text{сегм}}$, наибольшей $X_{\max}$ и средним значением $F_{\max}$;

— для сонорных звуков $t_{\text{сегм}}$, $X_{\max}$ и $F_{\max}$ принимают средние значения;

— для звонких звуков $t_{\text{сегм}}$ принимает среднее значение, $X_{\max}$ — низкое и $F_{\max}$ — низкое;

— для глухих звуков $X_{\max}$ принимает наименьшее значение, $F_{\max}$ принимает наибольшее значение;

— шипящие звуки характеризуются средним значением $t_{\text{сегм}}$, низким значением $X_{\max}$ и большим значением частоты $F_{\max}$.

В заключение представим результаты эксперимента для оценки точности классификации звуков. Для тестирования были записаны 100 других реализаций изолированно произнесенных звуков каждого класса для каждого диктора. В табл. 2 приведена точность распознавания для двух применяемых методов обучения.

Как видно из этой таблицы, точность классификации для алгоритмов обучения LERS и CART примерно совпадает. Можно заметить, что наиболее хорошо распознаются гласные звуки

■ ***Таблица 1.*** Наиболее характерные правила для каждого класса звуков с использованием алгоритмов LERS и CART

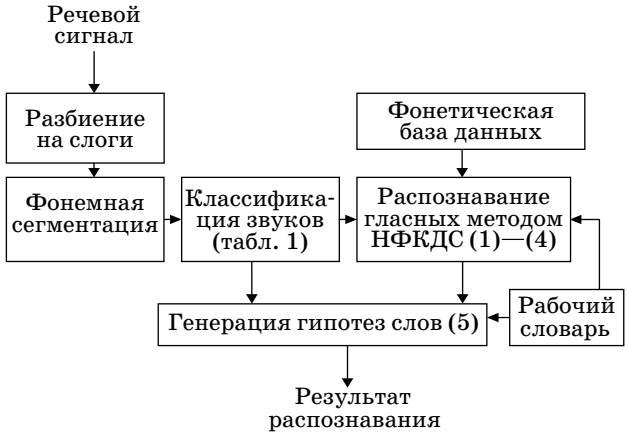
Класс	Правила	
	LERS	CART
Гласные	• $t_{\text{сегм}} \geq 6$; $0{,}75 \leq X_{\max} \leq 1$	• $4 \leq t_{\text{сегм}} \leq 8$; $0{,}9 \leq X_{\max} \leq 1$; $F_{\max} \geq 385$
Сонорные	• $t_{\text{сегм}} < 6$; $0{,}65 \leq X_{\max} \leq 1$ • $0{,}35 \leq X_{\max} \leq 0{,}65$; $265 \leq F_{\max} \leq 2210$	• $t_{\text{сегм}} < 9$; $0{,}5 \leq X_{\max} < 0{,}9$; $F_{\max} < 525$
Звонкие	• $0{,}20 \leq X_{\max} \leq 0{,}40$; $F_{\max} \leq 265$	• $0{,}19 \leq X_{\max} \leq 0{,}38$; $F_{\max} \leq 245$ • $t_{\text{сегм}} \geq 4$; $0{,}15 \leq X_{\max} < 0{,}19$; $F_{\max} < 215$
Глухие	• $0 < X_{\max} \leq 0{,}1$; $F_{\max} > 2210$	• $X_{\max} \leq 0{,}05$; $F_{\max} > 2140$
Шипящие	• $t_{\text{сегм}} < 5$; $0{,}1 \leq X_{\max} \leq 0{,}37$; $F_{\max} > 2210$	• $0{,}05 < X_{\max} \leq 0{,}25$; $F_{\max} > 1720$

■ ***Таблица 2.*** Точность распознавания различных классов звуков русского языка, %

Класс	Распознано		Не распознано		Распознано неверно	
	LERS	CART	LERS	CART	LERS	CART
Звонкие	76	73	12	14	12	13
Сонорные	77	74	15	15	8	11
Шипящие	73	72	7	8	20	20
Глухие	28	30	71	68	1	2
Гласные	96	96	2	3	2	1

(≈96 %), а хуже всех — глухие согласные (≈28 %), которые зачастую классифицируются как шум или пауза. Кажется, что такая точность классификации недостаточна для получения надежного решения, однако этот вывод требует уточнения. Действительно, точность классификации изолированных фонем обычно не превышает 40—60 % [6], так как для распознавания используются только фонетические и акустические особенности речевого сигнала. Идентификация фонем является лишь первым этапом при решении задачи распознавания слитной речи, точность решения которой значительно превышает точность распознавания составляющих входной речевой сигнал фонем за счет учета синтаксических, семантических и лексических особенностей языка [3]. В результате любое повышение точности распознавания фонем способно привести к резкому росту точности распознавания речи. Как будет показано далее в экспериментальном исследовании, выделенные признаки можно применять для повышения точности распознавания слов.

Структурная схема метода НФКДС, представленная на рис. 3, является расширением схемы пофонемного распознавания изолированных слов для метода ФДС [2]. Алгоритм распознавания изолированных слов, показанный на этом рисунке, отличается от традиционных методов пофонемного распознавания речи [16, 21] наличием блока распознавания гласных фонем методом НФКДС (1)—(4), а также процедурой (5) генерации гипотез.



■ ***Рис. 3.*** Структурная схема алгоритма распознавания изолированных слов методом НФКДС

Результаты экспериментальных исследований в задаче распознавания слов

В экспериментальной части работы проведем сравнение методов ФДС и НФКДС по точности распознавания изолированных слов русской речи для дикторозависимого и дикторонезависимого режимов. Сравнение сигналов проводилось на основе АР-модели в метрике Кульбака — Лейблера [7]:

$$\rho_{KL}\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)=\frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_r(f)}-\ln\frac{G_x(f)}{G_r(f)}\right)-1. \quad (6)$$

Здесь $G_x(f)$ — основанная на методе Берга [8] оценка СПМ входного сигнала $\mathbf{x}(t)$ как функция дискретной частоты f, а $G_r(f)$ — СПМ эталона r-й фонемы. Выражение (6) описывает алгоритм АРР на основе сопоставления СПМ в метрике Кульбака — Лейблера.

Важнейшее достоинство АР-модели в задачах АРР — это возможность нормировать речевые сигналы по дисперсии порождающих процессов: $\sigma_0^2=\sigma_x^2$, где σ_x^2 — дисперсия порождающего процесса. В работе [22] показано, что в таком случае $\rho\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)$ можно определить на основе выхода обеляющего фильтра (Whitening Filter — WF):

$$\rho_{WF}\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)=\frac{1}{2}\left[\frac{\sigma_r^2(\mathbf{x})}{\sigma_0^2}-1\right], \quad (7)$$

где $\sigma_r^2(\mathbf{x})$, $r=\overline{1,R}$ — выборочная оценка дисперсии отклика r-го обеляющего фильтра.

Также для сопоставления сигналов использовались мера близости Евклида с традиционными для АРР признаками MFCC (Mel Frequency Cepstral Coefficients) [3] и мера близости SD (Spectral Distortion) между АР-оценками СПМ, эквивалентная расстоянию между кепстральными коэффициентами [23]:

$$\rho_{SD}\left(\mathbf{x}(t)/\mathbf{x}_r^*\right)=\frac{1}{F}\sum_{f=1}^{F}\left(\ln\frac{G_x(f)}{G_r(f)}\right)^2. \quad (8)$$

Кроме того, в эксперименте для сопоставления точности распознавания гласных применялась современная система распознавания речи CMU Sphinx [24]. Настройка на диктора осуществлялась с помощью программы SphinxTrain на базе акустической модели русского языка из проекта ru4sphinx[1]. Система CMU Sphinx поддерживает

[1] https://github.com/zamiron/ru4sphinx/tree/master

■ ***Таблица 3.*** Оценка вероятности ошибки распознавания, % (словарь наименований лекарств)

Мера близости / система АРР	Признаки	Алгоритм классификации фонем	Дикторозависимый режим		Дикторонезависимый режим	
			ФДС	НФКДС	ФДС	НФКДС
Евклида	Кепстральные коэффициенты (MFCC)	Без признаков согласных	17,5±3	14±3	26±3,4	22±3
		LERS	14±2	11±2,2	**22**±3,2	19±3,1
		CART	15±3	12±1,8	24±3,1	20±4
Кульбака — Лейблера (6)	Оценка СПМ	Без признаков согласных	15±4	11,5±2	27±2,8	21,5±2
		LERS	**12,5**±2,3	**8**±2,4	24±2,7	**17**±2,4
		CART	13±3,2	9,5±2,3	25±2,1	22±1,9
Метод обеляющего фильтра (WF) (7)	Оценка АР-коэффициентов	Без признаков согласных	19±2,5	16±2,6	30±2,4	27,5±1,7
		LERS	16±2,5	12,5±3,2	27±2,3	25±2,4
		CART	16,5±2,6	12±3,4	26±1,8	24±2,2
Spectral Distortion (SD) (8)	Оценка СПМ	Без признаков согласных	18,5±3,2	16±2,8	29,5±3	27±2
		LERS	15,5±3,3	12±2	27±3,1	24±3,1
		CART	14,5±3,6	12±1,7	26,5±2,2	25±2,1
Pocketsphinx + ФДС	Кепстральные коэффициенты (MFCC)	–	14±8,2		24±7,1	
Google Voice Search	То же	–	–		32±5,2	

возможность автоматической адаптации словаря, поэтому она может быть использована в методе ФДС для распознавания гласных в слогах. Также в эксперименте было проведено сравнение с русскоязычной клиент-серверной версией системы Google Voice Search [25], качество распознавания которой считается для русского языка весьма высоким. В последнем случае все тестовые реализации слов были произнесены слитно (без выделения слогов).

Использовались следующие значения параметров алгоритма АРР: порядок АР-модели $p=15$, частота дискретизации $F=8$ кГц, $\Delta\mu=0{,}01$, ко всем тестовым реализациям слов/словосочетаний добавлялся белый гауссовый шум (отношение сигнал/шум 10 дБ). Параметр масштабирования α в методе НФКДС (2), (3) для каждого рассогласования подбирался экспериментально.

Фонетический алфавит $\left\{\mathbf{x}_r^*\right\}$, $r=\overline{1,R}$ был составлен из десяти изолированно произнесенных диктором гласных звуков русского языка. Рабочий словарь формировался автоматически из текстового файла, содержащего перечень лекарств из 1910 наименований. Для тестирования другой диктор (дикторонезависимый режим) или тот же диктор (дикторозависимый режим) произносил 3 раза все слова/словосочетания из словаря с четким выделением слогов.

Оценка вероятности ошибки распознавания, усредненная по шести дикторам (трем мужчинам и трем женщинам), представлена в табл. 3. Слово считается распознанным верно, если не существует ни одного слова с большей степенью принадлежности, чем истинное. Полужирным шрифтом в таблице выделены наилучшие результаты. В рассматриваемом эксперименте в ошибку распознавания включена ошибка алгоритма сегментации, т. е. неверное определение числа слогов в произнесенном слове. В данном случае ошибка сегментации оказалась равной 4 %.

Здесь метод НФКДС превосходит по точности распознавания метод ФДС на 3—7 % для различных режимов распознавания и для различных рассогласований. Так, например, для дикторозависимого режима и меры близости Кульбака — Лейблера с использованием алгоритма LERS ошибка распознавания для метода НФКДС равна 8 %, что на 4,5 % ниже аналогичного показателя для метода ФДС. Также можно заметить, что вероятность ошибки распознавания наиболее низкая для рассогласования Кульбака — Лейблера и традиционной метрики Евклида с MFCC-признаками. Автоматическое выделение признаков квазистационарных сегментов и их типизация (см. табл. 1) позволяют повысить точность распознавания на 2—6 % по сравнению с библиотекой Pocketsphinx для дикторозависимого режима, на 1—7 % для дикторонезависимого режима и на 4,5—24 % по сравнению с Google Voice Search для дикторозависимого и дикторонезависимого режимов. Между тем качество популярной системы Google Voice Search для словаря лекарств оказалось неудовлетворительным, так как большая часть наименований лекарствен-

■ ***Таблица 4.*** Оценка вероятности ошибки распознавания, % (словарь с наименованием городов)

Мера близости / система АРР	Признаки	Алгоритм классификации фонем	Дикторозависимый режим	
			ФДС	НФКДС
Евклида	Кепстральные коэффициенты (MFCC)	Без признаков согласных	25±4	21±2,8
		LERS	17±2	15±3,1
		CART	18±2,3	14,5±3,2
Кульбака — Лейблера (6)	Оценка СПМ	Без признаков согласных	22±4,2	18±1,8
		LERS	18±1,8	13,5±2,1
		CART	17,5±2,2	**13**±2,5
Метод обеляющего фильтра (WF) (7)	Оценка АР-коэффициентов	Без признаков согласных	27±2,1	24±2,3
		LERS	20,5±2,3	17±2,2
		CART	22±2,7	18±2,6
Spectral Distortion (SD) (8)	Оценка СПМ	Без признаков согласных	24,5±2,8	19±3,2
		LERS	18,5±2,6	15,5±3,4
		CART	19±3	15±2,2
Pocketsphinx + ФДС	Кепстральные коэффициенты (MFCC)	–	17,0±9,6	
Google Voice Search	То же	–	18±8	

ных препаратов отсутствует в универсальном словаре системы.

Вероятность ошибки распознавания для методов ФДС и НФКДС для словаря наименований 1920 городов РФ показана в табл. 4. Здесь приведены результаты только для дикторозависимого режима, точность которого значительно выше (см. табл. 3) по сравнению с дикторонезависимым режимом, при этом настройка под конкретного диктора занимает несколько минут [2, 4]. Ошибка сегментации составила 3 %.

Здесь вероятность ошибки распознавания без учета признаков достаточно велика (18—27 %), так как в таком случае из-за специфики словаря алгоритм АРР выдает много альтернативных решений. Использование признаков согласных позволяет понизить ошибку распознавания на 4,5—8 %. Также можно заметить, что предложенный алгоритм (см. рис. 3) превосходит систему распознавания CMU Sphinx на 1,5—4 %. Действительно, использование нечетких множеств и их пересечение (4) позволяют повысить точность классификации гласных и, как следствие, точность распознавания изолированных слов.

В этом эксперименте с распознаванием названий городов система дикторонезависимого распознавания Google Voice Search показала очень хорошие результаты, однако ее точность оказалась несколько ниже (на 1,5—5 %) точности предложенного подхода (см. рис. 3), что может быть объяснено требованием к послоговому произношению тестовых слов [2].

Заключение

В работе предложена модификация метода НФКДС, основанного на теории нечетких множеств, использующая классы согласных, которые могут быть выделены с помощью алгоритмов LERS или CART по выбранным нами акустическим признакам. Показано, что точность распознавания сонорных и звонких согласных достаточно велика, поэтому алгоритм классификации фонем может применяться для уточнения решения, полученного не только по методу НФКДС, но и по методу ФДС. Также экспериментально продемонстрировано, что метод НФКДС может применяться совместно с различными мерами близости и превосходит по точности распознавания слов традиционный метод ФДС на 3—7 %, библиотеку CMU Sphinx, применяемую для распознавания гласных в слоге методом ФДС, — на 1—7 % и современную систему APP Google Voice Search на 1,5—24 %.

Таким образом, увеличение степени принадлежности входного сигнала к эталонному слову за счет предложенного алгоритма классификации согласных звуков приводит к увеличению точности и к уменьшению количества альтернативных решений на выходе алгоритма распознавания. Одним из возможных направлений дальнейшего исследования метода НФКДС (2), (3) является его применение не только к гласным, но и к согласным звукам для их различения внутри одного типа (сонорные, шипящие и т. п.).

Литература

1. **Ронжин А. Л., Глазков С. В.** Метод автоматического распознавания голосовых команд и неречевых акустических событий // Информационно-управляющие системы. 2012. № 4. С. 74–77.
2. **Савченко А. В.** Адаптивный алгоритм распознавания речи на основе метода фонетического декодирования слов в задаче голосового управления // Информационные технологии. 2013. № 4. С. 34–39.
3. **Benesty J., Sondh M., Huang Y.** (eds.) Springer Handbook of Speech Recognition. – N. Y.: Springer, 2008. – 1159 p.
4. **Савченко В. В., Савченко А. В.** Метод фонетического декодирования слов в информационной метрике Кульбака — Лейблера для систем автоматического анализа и распознавания речи с повышенным быстродействием // Информационно-управляющие системы. 2013. № 2. С. 7–12.
5. **Козлов А. В., Саввина Г. В., Шелепов В. Ю.** Система пофонемного распознавания отдельно произносимых слов // Искусственный интеллект. 2003. № 1. С. 156–165.
6. **Савченко Л. В., Савченко А. В.** Алгоритм автоматического распознавания фонем на основе логики нечетких множеств в информационной метрике Кульбака — Лейблера // Вестник компьютерных и информационных технологий. 2013. № 3. С. 36–41.
7. **Kullback S.** Information Theory and statistics. – Dover Pub., 1997. – 399 p.
8. **Marple S. L.** (Jr.). Digital spectral analysis and with application, Englewood Cliffs. – New Jersey: Prentice-Hall, 1987. – 492 p.
9. **Hill J. E.** The Minimum of n Independent Normal Distributions. http://www.untruth.org/~josh/math/normalmin.pdf (дата обращения: 03.10.2013).
10. **Savchenko A. V., Savchenko L. V.** Fuzzy Phonetic Decoding Method in a Phoneme Recognition Problem // NOLISP-2013 Int.Conf, LNCS/LNAI 7911, Mons, Belgium, 2013. P. 176–183.
11. **Specht D. F.** Probabilistic neural networks// Neural Networks. 1990. Vol. 3. P. 109–118.
12. **Zadeh L. A.** Fuzzy Sets// Information Control. 1965. Vol. 8. P. 338–353.
13. **Матвеев Ю. Н.** Исследование информативности признаков речи для систем автоматической идентификации дикторов // Изв. вузов. Приборостроение. 2013. Т. 56. № 2. С. 47–51.
14. **Сорокин В. Н., Цыплихин А. И.** Сегментация и распознавание гласных// Информационные процессы. 2004. Т. 4. № 2. С. 202–220.
15. **Ронжин А. Л.** и др. Фонетико-морфологическая разметка речевых корпусов для распознавания и синтеза русской речи// Информационно-управляющие системы. 2006. № 6. С. 24–34.
16. **Дорохина Г. В.** Методы пофонемного распознавания, использующие свойства языка и речи// Искусственный интеллект. 2008. № 4. С. 332–338.
17. **Шанский Н. М., Иванов В. В.** Современный русский язык: в 3 ч. – М.: Просвещение, 1987. Ч. 1. – 192 с.
18. **Савченко Л. В., Акатьев Д. Ю.** Выделение признаков речевого сигнала на основе теории приближенных множеств в методе нечеткого фонетического декодирования слов// Нелинейная динамика в когнитивных исследованиях-2013: Всерос. конф., Н. Новгород, 24–27 сентября 2013 г. С. 148–151.
19. **Grzymala-Buss J. W.** A system for learning from examples based on rough sets (LERS)// Intelligent Decision Support: Handbook of Application and Advances of the Rough Sets Theory/ Slowinski R. (ed.) – Dordrecht: Kluwer Academic Publishers, 1992. – P. 3–18.
20. **Jordan M. I.** A statistical approach to decision tree modeling// Proc. of the Seventh Annual ACM Conf. of the Computation Learning Theory. N. Y.: ACM Press, 1994. P. 254–282.
21. **Ниценко А. В., Шелепов В. Ю.** Алгоритмы пофонемного распознавания слов наперед заданного словаря// Искусственный интеллект. 2004. № 4. С. 633–639.
22. **Савченко В. В., Акатьев Д. Ю., Карпов Н. В.** Автоматическое распознавание элементарных речевых единиц методом обеляющего фильтра// Изв. вузов. Радиоэлектроника. 2007. Вып. 4. С. 11–19.
23. **Wei B., Gibson J. D.** Comparison of distance measure in discrete spectral modeling // Proc. of IEEE 9th Digital Signal Processing Workshop. 2000. P. 1–4.
24. **Система** распознавания речи CMU Sphinx. http://cmusphinx.sourceforge.net/ (дата обращения: 28.10.2013).
25. **Schuster M.** Speech Recognition for Mobile Devices at Google// LNCS. 2010. Vol. 6230. P. 8–10.

UDC 004.934

Recognition Algorithm on the Basis of the Fuzzy Phonetic Coding-Decoding Method

Savchenko L. V.[a], Postgraduate Student, LyudmilaSavchenko@yandex.ru
[a]Nizhny Novgorod State Linguistic University, 31a, Minin St., 603155, Nizhny Novgorod, Russian Federation

Purpose: To increase accuracy of automatic recognition of Russian language in voice control applications. **Methods:** There has been proposed a modification of the fuzzy phonetic coding-decoding method which involves the known consonant features classified by means of machine learning algorithms on the basis of rough sets and decision trees. There have been demonstrated the most characteristic classification rules (IF..., THEN...) for each phoneme type. **Results:** There have been shown the experimental study results concerning

the problem of recognition of voice commands largely used in automatic speech processing for similarity measures (Kullback — Leibler information discrimination, Euclidean distance, Spectral distortion) and popular voice signal features (estimation of power spectral densities, coefficients of linear prediction, Mel frequency cepstral coefficients). It is shown that the accuracy of speech recognition for the proposed approach is 3—7%, 1—7%, 1.5—24% higher than the accuracy of the conventional phonetic word decoding, the modern speech recognition library CMU Sphinx and the popular Google Voice Search, respectively. **Practical relevance:** The increase of a degree of membership of an input signal to a reference word due to the proposed algorithm of consonant classification allows to increase the accuracy of recognition and to decrease an amount of output alternative words.

Keywords — Automatic Recognition of Isolated Words, Fuzzy Sets, Rough Sets, Decision Trees, Fuzzy Phonetic Coding-Decoding Method.

References

1. Ronzhin A. L., Glazkov S. V. The Method of Automatic Recognition of Voice Commands and Non-Speech Acoustic Events. *Informatsionno-upravliaiushchie sistemy*, 2012, no. 4, pp. 74–77 (In Russian).
2. Savchenko A. V. Adaptive Speech Recognition Algorithm on the Basis of the Words Phonetic Decoding Method in a Remote Control Problem. *Informatsionnye tekhnologii*, 2013, no. 4, pp. 34–39 (In Russian).
3. Benesty J., Sondh M., Huang Y. *Springer Handbook of Speech Recognition*. New York, Springer, 2008. 1159 p.
4. Savchenko V. V., Savchenko A. V. The Method of Words Phonetic Decoding Using Kullback — Leibler Information Discrimination for High-Speed Performance Systems of Automatic Speech Analysis and Recognition. *Informatsionno-upravliaiushchie sistemy*, 2013, no. 2, pp. 7–12 (In Russian).
5. Kozlov A. V., Savvina G. V., Shelepov V. Yu. Isolated Word Recognition System Based on Phoneme Recognition. *Iskusstvennyi intellekt*, 2003, no. 1, pp. 156–165 (In Russian).
6. Savchenko L. V., Savchenko A. V. Algorithm of Automatic Phoneme Recognition Based on the Fuzzy Sets Theory in the Kullback — Leibler Information Metric. *Vestnik komp'iuternykh i informatsionnykh tekhnologii*, 2013, no. 3, pp. 36–41 (In Russian).
7. Kullback S. *Information Theory and Statistics*. Dover Pub., 1997. 399 p.
8. Marple S. L. (Jr.). *Digital Spectral Analysis and with Application, Englewood Cliffs*. New Jersey, Prentice-Hall, 1987. 492 p.
9. Hill J. E. *The Minimum of n Independent Normal Distributions*, 2010. Available at: http://www.untruth.org/~josh/math/normalmin.pdf (accessed 3 October 2013).
10. Savchenko A. V., Savchenko L. V. Fuzzy Phonetic Decoding Method in a Phoneme Recognition Problem. *NOLISP-2013 Int. Conf., LNCS/LNAI 7911*. Mons, Belgium, 2013, pp. 176–183.
11. Specht D. F. Probabilistic Neural Networks. *Neural Networks*, 1990, vol. 3, pp. 109–118.
12. Zadeh L. A. Fuzzy Sets. *Information Control*, 1965, vol. 8, pp. 338–353.
13. Matveev Yu. Study of Informative Speech Features for Automatic Speaker Identification. *Izvestiia vuzov. Priborostroenie*, 2013, vol. 56, no. 2, pp. 47–51 (In Russian).
14. Sorokin V. N., Tsjplihin A. I. Segmentation and Recognition of Vowels. *Informatsionnye protsessy*, 2004, vol. 4, no. 2, pp. 202–220 (In Russian).
15. Ronzhin A. L., Karpov A. A., Lobanov B. M., Tsirulnik L. I., Jokisch O. Phonetic-Morphological Mapping of Speech Corpora for Recognition and Synthesis of Russian Speech. *Informatsionno-upravliaiushchie sistemy*, 2006, no. 6, pp. 24–34 (In Russian).
16. Dorohina G. V. Speech Recognition Methods Based on Phoneme Recognition that Use Features of Language and Speech, *Iskusstvennyi intellekt*, 2008, no. 4, pp. 332–338 (In Russian).
17. Shanskij N. M., Ivanov V. V. *Sovremennyi russkii iazyk* [Modern Russian Language]. Moskow, Prosveshchenie Publ., 1987. 192 p. (In Russian).
18. Savchenko L. V., Akatjev D. Yu. Features Action of Speech Signal which is Based on Rough Sets Theory in Fuzzy Phonetic Decoding Method. *Vserossiiskaia konferentsiia "Nelineinaia dinamika v kognitivnykh issledovaniiakh"* [All-Russian Conf. "Nonlinear Dynamic in Cognition Research"]. Nizhny Novgorod, 2013, pp. 148–151 (In Russian).
19. Grzymala-Buss J. W. A System for Learning from Examples Based on Rough Sets (LERS), In: *Intelligent Decision Support: Handbook of Application and Advances of the Rough Sets Theory*. Slowinski R. (ed.), Dordrecht, Kluwer Academic Publ., 1992, pp. 3–18.
20. Jordan M. I. A Statistical Approach to Decision Tree Modeling. *Proceeding of the Seventh Annual ACM Conference of the Computation Learning Theory*. New York, ACM Press, 1994, pp. 254–282.
21. Nitsenko A. V., Shelepov V. Yu. The Algorithms of the Phonetic Recognition of Isolated Words for a Given Dictionary. *Iskusstvennyi intellekt*, 2004, no. 4, pp. 633–639 (In Russian).
22. Savchenko V. V., Akatjev D. Yu., Karpov N. V. Automatic Recognition of the Minimal Speech Units by the Whitening Filter Method. *Izvestiia vuzov. Radioelektronika*, 2007, no. 4, pp. 11–19 (In Russian).
23. Wei B., Gibson J. D. Comparison of Distance Measure in Discrete Spectral Modeling. *Proceedings of IEEE 9th Digital Signal Processing Workshop*, 2000, pp. 1–4.
24. *The Speech Recognition System CMU Sphinx*. Available at: http://cmusphinx.sourceforge.net/ (accessed 28 October 2013).
25. Schuster M. Speech Recognition for Mobile Devices at Google. *LNCS*, 2010, vol. 6230, pp. 8–10.