РАДИОТЕХНИКА И ЭЛЕКТРОНИКА, 2019, том 64, № 3, с. 274–280

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ СИГНАЛОВ

УДК 004.934

НЕЧЕТКОЕ ФОНЕТИЧЕСКОЕ КОДИРОВАНИЕ РЕЧЕВЫХ СИГНАЛОВ В СИСТЕМАХ ОБРАБОТКИ ГОЛОСОВОЙ ИНФОРМАЦИИ

© 2019 г. Л. В. Савченко[1], А. В. Савченко[1, *]

[1]Национальный исследовательский университет Высшая школа экономики, Российская Федерация, 603155 Нижний Новгород, ул. Большая Печерская, 25/12

**E-mail: avsavchenko@hse.ru*

Поступила в редакцию 06.09.2017 г.
После доработки 25.04.2018 г.
Принята к публикации 14.05.2018 г.

Исследован фонетический подход для систем обработки голосовой информации. Разработан метод автоматического распознавания речевых сигналов, в котором каждому квазистационарному сегменту ставится в соответствие нечеткое множество фонем. Предложено использовать операцию вероятностной треугольной нормы для нечетких множеств, соответствующих входному фрейму и ближайшей к нему эталонной фонемы. Экспериментально показано, что разработанный метод позволяет на 1.5...5% снизить вероятность ошибочного распознавания по сравнению с известными аналогами.

DOI: 10.1134/S0033849419030173

ВВЕДЕНИЕ

Современные технологии трансформации речи в текст, основанные на скрытых марковских моделях (СММ) [1] и глубоких нейронных сетях [2], оказываются порой недостаточно эффективными [3] для многих систем обработки голосовой информации, например, систем голосового управления робототехникой широкого назначения [4]. Вычислительная сложность традиционных методов [1] затрудняет их применение в голосовом интерфейсе автономных систем, которые могут функционировать на малопроизводительном оборудовании. Кроме того, вероятность ошибочного распознавания сильно варьируется при наличии разнообразных акустических помех, акцента, дефектов речи, изменении физического и эмоционального состояния пользователя [5, 6]. Несомненный интерес здесь представляет пофонемная обработка речевых сигналов [7, 8], в которой появляется возможность выполнить быструю адаптацию на голос нового диктора и автоматическую настройку словаря [1, 9]. Перспективной реализацией пофонемного подхода является метод фонетического кодирования слов (ФКС) [3], позволяющий за счет использования принципов слоговой фонетики русского языка значительно снизить вероятность ложной тревоги [6]. Ключевой особенностью метода является предварительная редукция данных на основе принципа минимума информационного рассогласования Кульбака–Лейблера [10]: каждая фонема в фонетической базе данных (ФБД) задается с помощью центра кластера множества ее доступных реализаций [9].

Как известно, в алгоритмах пофонемной обработки речи, в том числе и в методе ФКС, каждый звук описывается собственной акустической моделью, при этом степень сходства различных звуков обычно не учитывается [3, 7, 11]. В результате на практике нередко требуется объединить модели близких по звучанию звуков в один кластер [6]. Такой подход приводит к значительному сокращению количества различимых звуков в ФБД и, как следствие, к увеличению числа альтернативных решений на выходе алгоритма распознавания [3]. Для преодоления указанной проблемы был предложен метод нечеткого фонетического кодирования (НФК) [4], основанный на представлении фонемы как нечеткого множества всех эталонных минимальных речевых единиц (МРЕ). В методе НФК для эталонных МРЕ и фреймов входного сигнала используются различные способы определения степеней принадлежности нечетких множеств, поэтому на практике в ряде случаев (например, при наличии помех) он оказывается недостаточно эффективным.

Цель данной статьи – найти способы повышения точности НФК для пофонемной обработки и распознавания голосовой информации на основе нечеткого множества эталонных фонем.

1. ФОНЕТИЧЕСКИЙ ПОДХОД В ЗАДАЧЕ РАСПОЗНАВАНИЯ ИЗОЛИРОВАННЫХ СЛОВ

Пусть задано множество из $L > 1$ слов, каждое из которых описывается в виде транскрипции – последовательности фонем из некоторого фонетического алфавита объема R. Задача состоит в том, чтобы поступившему на вход речевому сигналу X поставить в соответствие наиболее близкое к нему слово-эталон. Для сравнительной оценки эффективности различных методов решения задачи используется показатель вероятности ε ошибки распознавания, который обычно оценивается экспериментально в рамках перекрестной проверки (скользящего контроля) как среднее отношение некорректно распознанных слов к общему объему контрольной выборки. Кроме того, во многих работах применяется показатель $(1 - \varepsilon)$ – точность распознавания (доля правильно классифицированных слов) [1, 8].

В традиционных методах [1] решение сводится к автоматическому определению и последующему сопоставлению транскрипций произнесенного слова и всех эталонных слов. На первом этапе входной сигнал X разбивается на ряд непересекающихся квазистационарных (с неизменяющимися спектральными характеристиками [11, 12]) фреймов $\{\mathbf{x}(t)\}$, $t = \overline{1,T}$ длиной 10...30 мс (или $M = 80...240$ отсчетов при частоте дискретизации $F = 8000$ Гц), где T – общее число фреймов. Для каждого парциального сигнала – вектора-столбца отсчетов $\mathbf{x}(t)$ размерности M – вычисляются некоторые характерные признаки, например, кепстральные коэффициенты (MFCC, Mel Frequency Cepstral Coefficients). Вектор признаков каждой МРЕ (монофон или трифон) обычно описывается с помощью модели гауссовской смеси [1] или глубокой нейронной сети (ГНС) [2]. Для оценки параметров акустической модели используются речевые корпуса большого объема (десятки и даже сотни часов). На втором этапе с помощью методов динамического программирования (например, Dynamic Time Warping или СММ) выполняется динамическое выравнивание по темпу речи полученной последовательности признаков МРЕ и транскрипций слов из словаря. Слово, наиболее близкое в смысле метода максимального правдоподобия после временно́го выравнивания, и будет являться решением задачи.

Известны следующие основные проблемы таких традиционных методов [3]: высокая вычислительная сложность алгоритмов принятия решений, большой объем памяти для хранения акустической модели, сложность настройки на голос диктора. Для преодоления указанных недостатков может использоваться фонетический подход [6, 7], в котором каждая r-я фонема в ФБД задается множеством из $\{\mathbf{x}(t)\}$, $t = \overline{1,T}$ векторов отсчетов однофонемных эталонных реализаций $\mathbf{x}_{r,j}, j = \overline{1,J_r}$, где J_r – количество эталонных фонем r-го класса. В результате применения фонетического подхода появляется возможность выполнить быструю настройку на голос нового диктора, однако точность распознавания на практике оказывается ниже по сравнению с традиционными СММ-методами.

В работе [3] отмечено, что, в отличие от традиционных систем диктовки текстов, при построении голосового интерфейса во многих технических системах не требуется обработки сверхбольших словарей, а для повышения точности распознавания могут вводиться искусственные ограничения на стиль произнесения команд. Показано, что в таком случае альтернативой СММ-подходу может служить метод ФКС, в котором предварительно проводится редукция МРЕ [6] – одноименные реализации фонем r-го класса группируются вокруг информационного центра-эталона – речевой метки $\mathbf{x}_r^*$

$$\mathbf{x}_r^* = \underset{\mathbf{x}_{r,k}, k\in\{1,\ldots,J_r\}}{\operatorname{argmin}} \sum_{j=1}^{J_r} \rho_{KL}\left(\mathbf{x}_{r,k}, \mathbf{x}_{r,j}\right), \qquad (1)$$

которая характеризуется минимальной суммой информационных рассогласований Кульбака–Лейблера (KL) [10] между выборочными оценками автокорреляционных матриц $\mathbf{K}_{r,k}$ и $\mathbf{K}_{r,j}$ порядка p сигналов $\mathbf{x}_{r,k}$ и $\mathbf{x}_{r,j}$ соответственно:

$$\rho_{KL}\left(\mathbf{x}_{r,k}, \mathbf{x}_{r,j}\right) = \\ = \left(\frac{1}{2}\ln\frac{\left|\mathbf{K}_{r,k}\right|}{\left|\mathbf{K}_{r,j}\right|} + \frac{1}{2}\operatorname{tr}\left(\mathbf{K}_{r,j}\left(\mathbf{K}_{r,k}\right)^{-1}\right) - \frac{p}{2}\right), \qquad (2)$$

где использованы обозначения $|\mathbf{K}|$ и $\operatorname{tr}(\mathbf{K})$ для определителя и следа матрицы, соответственно.

При использовании традиционной для задач обработки речи [1] авторегрессионной (АР) модели (или модели линейного предсказания речи [11, 12]) порядка p (обычно $p = 12...20$) последнее выражение с точностью до постоянного множителя эквивалентно дивергенции Итакуры–Саито [1, 13, 14] между АР-оценками спектральной плотности мощности (СПМ) $G_{r,k}(f)$ и $G_{r,j}(f), f \in \{1, 2, \ldots, F\}$ эталонов $\mathbf{x}_{r,k}$ и $\mathbf{x}_{r,j}$:

$$\rho_{KL}\left(\mathbf{x}_{r,k}, \mathbf{x}_{r,j}\right) = \frac{2}{F}\sum_{f=1}^{F/2}\left(\frac{G_{r,k}(f)}{G_{r,j}(f)} - \ln\frac{G_{r,k}(f)}{G_{r,j}(f)} - 1\right). \qquad (3)$$

Нередко близкие по звучанию звуки объединяются в один кластер: r-й фонеме ставится в соответствие фонетический код – номер кластера $c(r) \in \{1,\ldots,C\}$, где $C \le R$ – число различных кластеров.

В процессе распознавания во входном слове X с помощью амплитудного детектора выделяется N слогов. Границы n-го слога ($n = \overline{1,N}$) определяются с точностью до номера фрейма. Далее каждому слогу ставится в соответствие один из R эталонов $\mathbf{x}^*_{\nu(t)}$ из ФБД с помощью известной реализации максимально правдоподобного решения задачи проверки гипотез об автокорреляционных матрицах гауссовского процесса [6] на основе принципа минимума информационного рассогласования Кульбака–Лейблера [10]:

$$v(t) = \arg\min_{r=\overline{1,R}} \rho_{KL}\left(\mathbf{x}(t), \mathbf{x}^*_r\right), \quad t = \overline{1,T}, \tag{4}$$

после чего с помощью простого голосования осуществляется агрегация решений (4) для каждого фрейма n-го слога. Итоговое решение принимается в пользу слова из словаря, последовательность фонем в котором наиболее близка фонемам, распознанным во входном сигнале.

2. НЕЧЕТКОЕ ФОНЕТИЧЕСКОЕ КОДИРОВАНИЕ РЕЧЕВЫХ СИГНАЛОВ

В работе [4] выделен основной недостаток практической реализации метода ФКС – из-за близости многих звуков число фонем R зачастую намного превышает число используемых при распознавании фонетических кодов C, что приводит к наличию большого числа альтернативных слов на выходе алгоритма [3]. Для устранения этого недостатка для задачи распознавания фонем был предложен метод НФК [4, 15], в котором используется модель фонемы, основанная на теории нечетких множеств [16]. В ней каждый i-й звук описывается с помощью нечеткого множества МРЕ $\left\{\left(\mathbf{x}^*_r, \mu_i\left(\mathbf{x}^*_r\right)\right)\right\}$, а не только одного информационного центра-эталона $\mathbf{x}^*_i$. Степень принадлежности $\mu_i\left(\mathbf{x}^*_r\right)$ нечеткого множества определяется как условная вероятность $P\left(\mathbf{x}^*_r \middle| \mathbf{x}^*_i\right)$ принятия решения в пользу эталона $\mathbf{x}^*_r$ (4) при распознавании i-ой фонемы. Оценка этой вероятности может быть получена [3] на основе известного асимптотического распределения рассогласования Кульбака–Лейблера [10]:

$$\mu_i\left(\mathbf{x}^*_r\right) \overset{def}{=} \hat{P}\left(\mathbf{x}^*_r \middle| \mathbf{x}^*_i\right) = \frac{1}{\sqrt{2\pi}} \times \times \int_{-\infty}^{+\infty} \left(\exp\left(-\frac{t^2}{2}\right) \prod_{k=1, k\neq i}^{R} \left(\frac{1}{2} - \frac{1}{2}\Phi\left(\tilde{t}_{r,i,k}(t)\right)\right)\right) dt, \tag{5}$$

где

$$\tilde{t}_{r,i,k}(t) = \frac{t\sqrt{2\lambda\rho_{KL}(\mathbf{x}^*_r, \mathbf{x}^*_i) + p(p+1)/4} + \lambda\left(\rho_{KL}(\mathbf{x}^*_r, \mathbf{x}^*_i) - \rho_{KL}(\mathbf{x}^*_k, \mathbf{x}^*_i)\right)}{\sqrt{2\lambda\rho_{KL}(\mathbf{x}^*_k, \mathbf{x}^*_i) + p(p+1)/4}},$$

$\Phi(\tilde{t})$ – функция Лапласа, $\lambda = 4\pi(M-p)/p$ – параметр масштабирования.

Каждому фрейму входного сигнала $\mathbf{x}(t)$ также ставится в соответствие нечеткое множество вида $\left\{\left(\mathbf{x}^*_r, \mu\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right)\right)\right\}$. Степень принадлежности $\mu\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right)$ определяется как апостериорная вероятность $P\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right)$ принадлежности фрейма $\mathbf{x}(t)$ к r-й гласной, которая для критерия минимума информационного рассогласования (4) может быть оценена как [3]:

$$\mu\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right) \overset{def}{=} \hat{P}\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right) = = \frac{\exp\left(-\lambda\rho_{KL}\left(\mathbf{x}(t), \mathbf{x}^*_r\right)\right) p_r}{\sum_{k=1}^{R} \exp\left(-\lambda\rho_{KL}\left(\mathbf{x}(t), \mathbf{x}^*_k\right)\right) p_k}. \tag{6}$$

Здесь $p_r, r = \overline{1,R}$ – априорная вероятность появления r-го класса гласных звуков (частота появления фонемы в языке). Для повышения точности распознавания фонем используется операция пересечения нечетких множеств $\left\{\left(\mathbf{x}^*_r, \mu\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right)\right)\right\}$ и $\left\{\left(\mathbf{x}^*_r, \mu_{\nu(t)}\left(\mathbf{x}^*_r\right)\right)\right\}$, где $\nu(t)$ – номер ближайшего к t-му фрейму эталона из ФБД (4):

$$\mu(r,t) = \max\left(\mu_{\nu(t)}\left(\mathbf{x}^*_r\right), \mu\left(\mathbf{x}^*_r \middle| \mathbf{x}(t)\right)\right). \tag{7}$$

Для рассогласования Кульбака–Лейблера это выражение приводит к существенному снижению степеней принадлежности в случае ошибки распознавания [4]. В отличие от метода ФКС, агрегация решений для каждого слога в НФК осуществляется не простым голосованием, а усреднением степеней принадлежности (7) для всех фреймов этого слога.

3. ПРЕДЛОЖЕННЫЙ АЛГОРИТМ

При практической реализации метода НФК использование оценки условной вероятности перепутывания фонем (5) для пользовательской ФБД оказывается некорректным в связи с известной вариативностью речи [1]. Рассмотрим пример: при $i = r$ рассогласование $\rho_{KL}(\mathbf{x}_r^*, \mathbf{x}_r^*) = 0$, однако в действительности расстояние между одноименными реализациями фонемы может быть весьма велико, поэтому оценка (5) вероятности правильного распознавания r-й фонемы $P\left(\mathbf{x}_r^* \middle| \mathbf{x}_r^*\right)$ окажется завышенной. А добавление к эталонным МРЕ аддитивного шума, предложенное в работе [3], приводит к значительным различиям при определении нечетких множеств для эталонных фонем (5) и входных фреймов, что, в свою очередь, сказывается на точности итогового решения. Поэтому в данной работе для определения степени принадлежности эталонной МРЕ предлагается использовать оценку апостериорной вероятности, аналогичную (6):

$$\mu_i\left(\mathbf{x}_r^*\right) = \frac{\exp\left(-\lambda\rho_{KL}\left(\mathbf{x}_i^*, \mathbf{x}_r^*\right)\right) p_r}{\sum\limits_{k=1}^{R} \exp\left(-\lambda\rho_{KL}\left(\mathbf{x}_i^*, \mathbf{x}_k^*\right)\right) p_k}. \qquad (8)$$

С учетом того, что степени принадлежности в нечетких множествах определяются как дискретные распределения (6), (8), для их комбинирования вместо операции максимума (7) будем использовать вероятностную треугольную норму (вероятностное пересечение или алгебраическое произведение):

$$\mu(r,t) = \mu_{\nu(t)}\left(\mathbf{x}_r^*\right)\mu\left(\mathbf{x}_r^* \middle| \mathbf{x}(t)\right). \qquad (9)$$

Отметим, что обычно доступно более одной эталонной реализации МРЕ для каждой фонемы. Тогда вначале аналогично (8) оценим апостериорную вероятность для каждой j-й реализации i-й фонемы:

$$P\left(\mathbf{x}_r^* \middle| \mathbf{x}_{i,j}\right) = \frac{\exp\left(-\lambda\rho_{KL}\left(\mathbf{x}_{i,j}, \mathbf{x}_r^*\right)\right) p_r}{\sum\limits_{k=1}^{R} \exp\left(-\lambda\rho_{KL}\left(\mathbf{x}_{i,j}, \mathbf{x}_k^*\right)\right) p_k}, \qquad (10)$$
$$j = \overline{1, J_i}.$$

Заметим, что в случае, если хотя бы одна из реализаций истинной фонемы близка к эталону $\mathbf{x}_{\nu(t)}^*$, но при этом рассогласование между информационными центрами-эталонами $\rho_{KL}\left(\mathbf{x}_i^*, \mathbf{x}_r^*\right)$ достаточно велико, то оценка близости r-го эталона и t-го фрейма $\mu(r,t)$ (9) может оказаться слишком малой. Поэтому в данной работе для ее повышения предлагается определить степень принадлежности $\mu_i\left(\mathbf{x}_r^*\right)$ с помощью операции вероятностного объединения (или алгебраической суммы):

$$\mu_i\left(\mathbf{x}_r^*\right) = 1 - \prod_{j=1}^{J_i}\left(1 - P\left(\mathbf{x}_r^* \middle| \mathbf{x}_{i,j}\right)\right). \qquad (11)$$

Ниже представлен предлагаемый алгоритм распознавания фонем на основе нечеткого фонетического кодирования речевых сигналов:

1) для каждой фонемы среди множества ее реализаций выбирается информационный центр-эталон (1),

2) оцениваются апостериорные вероятности (10),

3) вычисляются степени принадлежности каждого r-го информационного центра к i-й фонеме (11);

4) входной речевой сигнал разбивается на T фреймов фиксированной длительности,

5) для вектора из M отсчетов каждого фрейма определяется код ближайшего информационного центра из ФБД (4),

6) оценивается степень принадлежности к каждому эталону из ФБД (7),

7) выполняется операция вероятностного пересечения (9),

8) выполняется агрегация (усреднение) нечетких множеств всех фреймов,

9) итоговое решение принимается в пользу звука с наибольшей средней степенью принадлежности.

Вычислительная сложность такого алгоритма асимптотически совпадает со сложностью реализации метода ФКС и может быть оценена как $O(TF(p + R))$. Затраты памяти для хранения акустической модели составляют $O((F + R)R)$, так как для вычисления рассогласования Кульбака–Лейблера для каждой фонемы требуется сохранить $F/2$ значений СПМ. Как известно, важнейшей особенностью рассогласования Кульбака–Лейблера в (1) является возможность его адаптивной реализации на основе метода обеляющего фильтра [1, 3, 6], что позволяет снизить вычислительную сложность примерно в F/M раз – до $O(TRp(M-p))$. Сложность по затратам памяти также снижается до $O((p + R)R)$, так как для каждого обеляющего фильтра сохраняются только p коэффициентов линейного предсказания.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Рассмотрим результаты сравнительного анализа предложенного подхода (1), (4), (8)–(11) с методом ФКС, оригинальным НФК (1)–(7) и традиционными способами распознавания речи.

Таблица 1. Результаты распознавания изолированных звуков для русского языка

Метод	$J_r = 1$		$J_r = 4$	
	вероятность ошибки ε	доверительный интервал	вероятность ошибки ε	доверительный интервал
Система “Профессор Хиггинс”	0.650 ± 0.048	[0.626; 0.674]	0.427 ± 0.035	[0.402; 0.452]
ФКС	0.667 ± 0.045	[0.644; 0.69]	0.493 ± 0.038	[0.468; 0.518]
НФК (1)–(7)	0.672 ± 0.051	[0.649; 0.695]	0.439 ± 0.042	[0.414; 0,464]
Предложенный подход (1), (4), (8)–(11), сопоставление СПМ	0.617 ± 0.04	[0.593; 0.641]	0.367 ± 0.032	[0.343; 0.391]
Предложенный подход (1), (4), (8)–(11), сопоставление признаков из системы “Профессор Хиггинс”	0.567 ± 0.046	[0.542; 0.592]	0.313 ± 0.026	[0.290; 0.336]

Таблица 2. Результаты распознавания изолированных звуков для английского языка

Метод	$J_r = 1$		$J_r = 4$	
	вероятность ошибки ε	доверительный интервал	вероятность ошибки ε	доверительный интервал
Система “Профессор Хиггинс”	0.593 ± 0.045	[0.566; 0.62]	0.327 ± 0.030	[0.301; 0.353]
ФКС	0.643 ± 0.048	[0.617; 0.669]	0.333 ± 0.032	[0.307; 0.359]
НФК (1)–(7)	0.651 ± 0.050	[0.625; 0.677]	0.315 ± 0.035	[0.29; 0.34]
Предложенный подход (1), (4), (8)–(11), сопоставление СПМ	0.600 ± 0.038	[0.573; 0.627]	0.267 ± 0.028	[0.243; 0.291]
Предложенный подход (1), (4), (8)–(11), сопоставление признаков из системы “Профессор Хиггинс”	0.533 ± 0.042	[0.506; 0.56]	0.220 ± 0.024	[0.197; 0.243]

Запись речевого сигнала осуществлялась через внешний микрофон с функцией шумоподавления из гарнитуры A4Tech HS в формате моно, частота дискретизации $F = 8000$ Гц, разрядность квантования 16 бит на отсчет. Из сигнала были удалены начальные и конечные паузы. Порядок АР-модели $p = 20$, отношение сигнал/шум (С/Ш) 30 дБ, количество отсчетов в одном фрейме $M = 120$ (15 мс).

В первом эксперименте рассматривалась задача распознавания изолированно произнесенных звуков русского и английского языка. Эта задача имеет важное практическое значения для систем постановки произношения, поэтому результаты предложенного подхода сопоставлялись с известной системой обучения речи серии “Профессор Хиггинс” компании “ИстраСофт”, в которых фрейм речевого сигнала описывается с помощью дисперсий выходов полосовых фильтров, настроенных на девять различных диапазонов частот [17]. Тестирование проводилось группой из пяти пользователей (двух мужчин и трех женщин), каждый из которых 50 раз произносил каждый звук. Таким образом, общее число испытаний для каждого диктора составило 1550 для русского языка и 1300 для английского языков. Для наполнения ФБД при дикторонезависимого режима было использовано по одной ($J_r = 1$) реализации каждой фонем из программы “Профессор Хиггинс”, произнесенных эталонным пользователем. Кроме того, для тестирования дикторозависимого распознавания в ФБД добавлялись по три реализации фонем пользователя, в таком случае для каждого звука было доступно $J_r = 4$ эталонных сигналов.

В табл. 1 и 2 представлены результаты сравнительного анализа вероятности ε ошибки распознавания изолированных звуков в формате среднее (по пользователям) ± среднеквадратичное отклонение и доверительный интервал для вероятности ε, вычисленный при фиксированной доверительной вероятности 0.95 [18].

Здесь применение подхода к обучению, в котором пользователь может обучаться не только на фонемах эталонного диктора, но и на лучших из произнесенных им ранее звуков ($J_r = 4$), позволило увеличить точность оценки качества обучения на 18...25% для русского языка и на 25...33% для английского. Предложенная в данной работе модификация метода НФК обладает наибольшей точностью: на 5...10% выше по сравнению с си-

Таблица 3. Результаты распознавания изолированных слов/фраз

Метод	"Населенные пункты"		"Лекарства"	
	вероятность ошибки ε	доверительный интервал	вероятность ошибки ε	доверительный интервал
СММ	0.12 ± 0.042	[0.109; 0.131]	0.14 ± 0.045	[0.129; 0.151]
ФКС	0.10 ± 0.041	[0.09; 0.11]	0.12 ± 0.040	[0.11; 0.13]
НФК (1)–(7)	0.09 ± 0.044	[0.081; 0.099]	0.10 ± 0.042	[0.09; 0.11]
Предложенный подход (1), (4), (8)–(11)	0.065 ± 0.036	[0.057; 0.073]	0.08 ± 0.038	[0.071; 0.089]

стемой "Профессор Хиггинс". Кроме того, разработанный алгоритм оказался на 6...15% точнее по сравнению с оригинальными методами ФКС и НФК. Несмотря на то, что в ряде случаев доверительный интервал вероятности ошибки предложенного подхода пересекается с доверительными интервалами известных методов, критерий МакНемара [19] на уровне значимости 0.95 показал, что различия в точности разработанного метода с аналогами во всех случаях являются значимыми. Более того, эксперимент показал, что предложенный подход может быть реализован не только для рассогласования Кульбака–Лейблера между АР-оценками СПМ, но и с другими, специально подобранными для конкретной задачи признаками речевого сигнала.

Во втором эксперименте рассмотрена задача распознавания изолированных слов русского языка. Основным требованием к произношению было разделение слов на слоги с четкой паузой между ними (не менее 120 мс) [3, 20]. В качестве ФБД использовались все гласные звуки русского языка. Звуки /а/ и /й/ /а/, /э/ и /й/ /э/, /о/ и /й/ /о/, /и/ и /ы/, /у/ и /й/ /у/ в методе ФКС были объединены в $C = 5$ кластеров. В режиме настройки диктор в течение трех раз четко проговаривал каждый из десяти гласных звуков.

При тестировании использовались два набора фраз [4, 20]:

а) названия 1830 крупных населенных пунктов России (вместе с содержащими их областями, например, "Кстово (Нижегородская)" (далее – "Населенные пункты").

б) список из 1913 названий лекарств, продаваемых в одной из аптек Н. Новгорода (далее – "Лекарства").

Каждый диктор в идеальных условиях произнес по две реализации каждой фразы из указанных наборов. Таким образом, общее число испытаний составило 3660 и 3826 для "Населенных пунктов" и "Лекарств" соответственно.

Для распознавания гласных фонем в методе ФКС наряду с сопоставлением фреймов с эталонами из ФБД в метрике Кульбака–Лейблера применялась традиционная реализация СММ из библиотеки CMU Pocketsphinx. В табл. 3 представлены оценки вероятности ε ошибки распознавания и их доверительные интервалы для доверительной вероятности 0.95.

Для тестирования устойчивости методов распознавания речи к наличию помех к каждому записанному речевому сигналу искусственно добавлялся аддитивный гауссовский шум [20, 21]. На рис. 1 и рис. 2 приведены зависимости вероятности ошибки ε от отношения С/Ш для каждого набора данных.

Из анализа рисунков видно, что, во-первых, предложенный подход на основе операций с нечеткими множествами (8)–(11) превосходит по точности распознавания традиционные СММ-методы на 6...11% и базовый метод ФКС на 3...6%, при этом различия в точности являются статистически значимыми. При этом среднее время распознавания и сложность по затратам памяти оказываются на порядок меньше по сравнению с СММ-подходом. Во-вторых, использование в предложенной модификации НФК дополнительной информации обо всех эталонных реализациях при построении нечеткого множества фонем (10), (11) и переход от треугольной нормы Заде (7) к вероятностному пересечению позволили на 2.5...10% снизить вероятность ошибки оригинального метода (1)–(7). При этом различия в НФК и разработанной модификации особенно заметны при наличии достаточно сильных помех (10...15 дБ), в результате которых оценка условной вероятности (5) становится неточной. В-третьих, точность распознавания СММ-методов и оригинального НФК при сильных помехах (отношение С/Ш 10 дБ) снижается на 14...15% и 16...17.5% соответственно. В то же время для предложенного подхода вероятность ошибки при добавлении таких помех повышается на 8...10%, т.е. разработанная модификация НФК является более помехоустойчивой по сравнению с оригинальным методом.

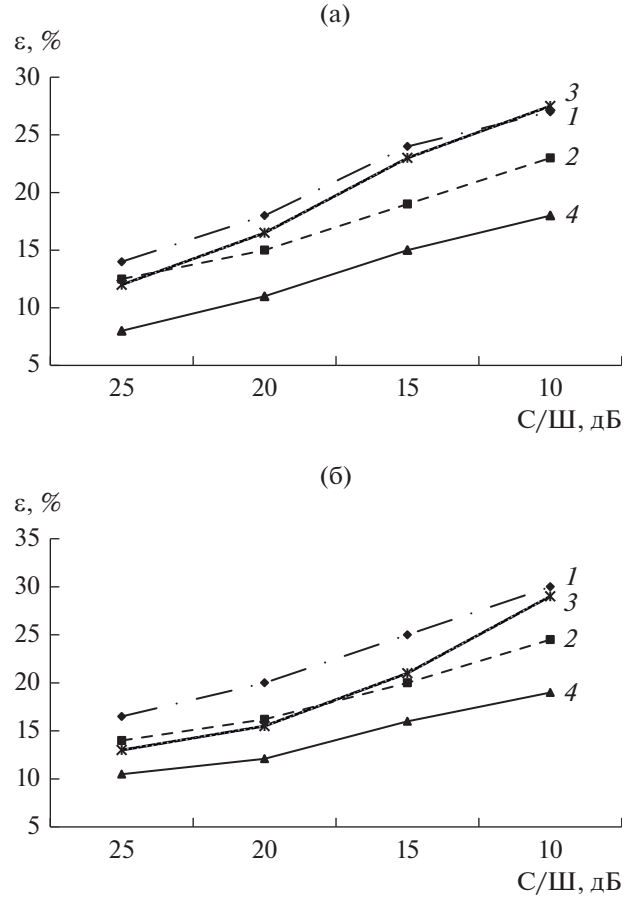

Рис. 1. Оценки вероятности ошибки распознавания изолированных слов ε в зависимости от отношения сигнал/шум (С/Ш) для наборов фраз "Населенные пункты" (а) и "Лекарства" (б): кривая *1* – СММ; кривая *2* – ФКС; кривая *3* – НФК (1)–(7); кривая *4* – предложенный подход (1), (4), (8)–(11).

ЗАКЛЮЧЕНИЕ

Применение теории нечетких множеств в задачах автоматического распознавания речи обычно связывают с формированием акустической модели каждой фонемы как нечеткого множества признаков [7, 11]. В противоположность такому подходу в методе НФК [4] предложено рассматривать ФБД как совокупность нечетких множеств эталонных фонем. В данной статье представлен способ практической реализации метода НФК, позволяющей получать правильные решения даже при наличии помех (рис. 1, 2) за счет применения более согласованного с (6) определения степени принадлежности эталонной фонемы (8), а также использования в выражениях (10), (11) более полной информации о доступных эталонных МРЕ. Преимущества разработанного алгоритма по сравнению с традиционными СММ-методами обработки и распознавания речевых сигналов будут особенно заметны при использовании на малопроизводительном оборудовании в автономном режиме (без доступа к сети Интернет) [21]. Действительно, предложенный подход (равно как и базовые методы ФДС и НФК) не требует большого объема памяти для хранения акустической модели и позволяет распознавать голосовые команды к режиме квази-реального времени.

Статья подготовлена в результате проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета "Высшая школа экономики" (НИУ ВШЭ).

СПИСОК ЛИТЕРАТУРЫ

1. *Springer Handbook of Speech Recognition* / Eds. Benesty J., Sondh M., Huang Y. N.Y.: Springer, 2008.
2. *Кипяткова И.С., Карпов А.А.* // Автоматика и телемеханика. 2017. № 5. С. 110.
3. *Савченко А.В.* // РЭ. 2014. Т. 59. № 4. С. 339.
4. *Savchenko A.V., Savchenko L.V.* // Pattern Recognition Lett. 2015. № 65. P. 145.
5. *Корсун О.Н., Финаев И.М., Чучупал В.Я., Яцко А.А.* // Наука и образование. 2013. № 1. С. 103.
6. *Савченко В.В.* // РЭ. 2017. Т. 62. № 7. С. 681.
7. *Каргин А.А., Шарий Т.В.* // Искусственный интеллект. 2010. № 3. С. 210.
8. *Savchenko A.V.* Search Techniques in Intelligent Classification Systems. Switzerland: Springer, 2016.
9. *Савченко В.В.* // РЭ. 2017. Т. 62. № 7. С. 681.
10. *Савченко А.В.* // РЭ. 2016. Т. 61. № 4. С. 373.
11. *Kullback S.* Information Theory and Statistics. N.Y.: Dover Publications, 1997.
12. *Ramou N., Guerti M.* // J. Commun. Technol. Electron. 2014. V. 59. № 11. P. 1274.
13. *Анциперов В.Е.* // РЭ. 2008. Т. 53. № 1. С. 73.
14. *Савченко В.В.* // РЭ. 1997. Т. 42. № 4. С. 426.
15. *Gray R.M., Gray A.H., Masuyama Y.* // IEEE Trans. 1980. V. ASSP-28. № 4. P. 367.
16. *Savchenko L.V., Savchenko A.V.* // Proc. Int. Conf. on Nonlinear Speech Processing (NOLISP2013) Mons. 19–21 June 2013. Lecture Notes in Artificial Intelligence. Switzerland: Springer, 2013. V. 7911. P. 176.
17. *Halavati R., Shouraki S.B., Zadeh S.H.* // Appl. Soft Comp. 2007. V. 7. № 3. P. 828.
18. ЗАО "ИстраСофт". Профессор Хиггинс: Английский без акцента! // Свид-во о гос. регистрации программы для ЭВМ № 2009614030. Опубл. 30.07.2009.
19. *Савченко В.В.* // Науч. ведомости Белгород. гос. ун-та. Серия: Экономика. Информатика. 2015. Т. 33. № 1. С. 74.
20. *Gillick L., Cox S.J.* // Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP-89). 1989. V. 1. P. 532.
21. *Savchenko A.V.* // Automation and Remote Control. 2013. V. 74. № 7. P. 1225.
22. *Savchenko A.V.* // Proc. Int. Joint Conf. on Rough Sets (IJCRS 2017) Olzstyn. 02–07 Jul. 2017. Lecture Notes in Computer Sci. Switzerland: Springer, 2017. V. 10314. P. 264.