УДК 621.372:519.72

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ МЕТОДИКИ ФОРМИРОВАНИЯ ФОНЕТИЧЕСКОЙ БАЗЫ ДАННЫХ ДИКТОРА ИЗ НЕПРЕРЫВНОГО ПОТОКА ЕГО РАЗГОВОРНОЙ РЕЧИ

В. В. Савченко,
доктор техн. наук, профессор
Д. Ю. Акатьев,
канд. техн. наук, доцент
Нижегородский государственный лингвистический университет

Ставится задача автоматического формирования фонетической базы данных диктора из непрерывного потока его устной речи. Предложена методика ее решения на основе когнитивной акустической модели минимальных звуковых единиц типа фонетического кластера в информационной метрике Кульбака — Лейблера. Рассмотрен пример практической реализации методики, представлены программа и результаты экспериментальных исследований.

Ключевые слова *— речь, русская речь, фонема, фонетический анализ речи, критерий минимума информационного рассогласования.*

Введение

При анализе разговорной (устной) речи на русском языке мы опираемся на наши точные знания в отношении его фонетического строя, количественного и качественного состава используемой фонетической системы, а также закономерностей ее функционирования в разговорной речи. Этими знаниями мы пользуемся, например, при транскрибировании потока речи. Однако при анализе разговорной речи на неизвестном языке нам недоступна, в общем случае, информация, относящаяся к его фонетической структуре. Тогда мы можем либо, опираясь на наш лингвистический опыт, давать участкам речевого потока приблизительную интерпретацию в рамках международного фонетического алфавита, либо, обратившись к акустическим понятиям, членить речь на некие минимальные звуковые единицы (МЗЕ) с соответствующими метками. Очевидно, что второй подход, положенный в основу информационной теории восприятия речи и ее когнитивной кластерной модели МЗЕ [1], со всех точек зрения наиболее информативен и универсален. Множество меток всех МЗЕ и составит в итоге звуковой строй данного диалекта (или языка) или его фонетическую базу данных (ФБД). Ее решению с использованием нового математического аппарата информационной теории восприятия речи и посвящена настоящая статья.

Краткие теоретические сведения

Фонема — это нечленимая, т. е. элементарная (минимальная) речевая единица (ЭРЕ). Несмотря на существующие различия в реализациях $\mathbf{x}_{r,j},\ j=\overline{1,J_r},\ J_r>>1$, некоторой r-й фонемы, все они воспринимаются человеком как нечто общее, иначе речь утратила бы свою информативность. Можно поэтому утверждать, что одноименные реализации в сознании человека группируются в соответствующие классы или речевые образы фонем $X_r=\left\{\mathbf{x}_{r,j}\right\},\ r=\overline{1,R}$, вокруг некоторого центра — эталонной метки данного образа. В информационной теории восприятия речи указанные эталоны определяются в строгом теоретико-информационном смысле: речевая метка $\mathbf{x}_r^*\subset X_r$ образует информационный центр-эталон r-го речевого образа, если в пределах множества X_r она характеризуется минимальной суммой информационных рассогласований по Кульбаку — Лейблеру относительно всех других его меток-реализаций $\mathbf{x}_{r,j},\ j=\overline{1,J_r}$.

Нетрудно увидеть, что именно в понятии информационного центра r-го множества реализаций одноименных МЗЕ X_r дается наиболее информативное описание свойств соответствующей фонемы. А множество всех информационных центров $\{\mathbf{x}_r^*\}$ определяет понятие ФБД для данного диктора. Одновременно становится очевидным и механизм формирования самого этого множества. Сначала анализируемый (входной) речевой сигнал $X(t)$ в дискретном времени $t = 0, 1, ...$ разбивается на ряд последовательных сегментов данных $x(t)$ длиной в одну МЗЕ — примерно 10–15 мс. После этого каждый такой парциальный сигнал рассматривается в пределах конечного списка фонем $\{X_r\}$ и отождествляется с той X_ν из них, которая отвечает критерию минимума информационного рассогласования (МИР) относительно сигнала $x(t)$. Это известная формулировка критерия МИР в задачах автоматического распознавания речи. Задача существенно упрощается, если воспользоваться гауссовой (нормальной) аппроксимацией закона распределения каждой фонемы вида $\mathbf{P}_r = N(\mathbf{K}_r)$, где $\mathbf{K}_r$ — автокорреляционная матрица размера $n \times n$, $n \geq 1$.

Выделим в анализируемом речевом сигнале $X(t)$ от некоторого диктора первые L отсчетов из соображений сохранения в них свойства приблизительной стационарности или однородности распределения $\mathbf{P}_r$. Например, при стандартной частоте дискретизации телефонного канала связи 8 кГц обычно полагают $L = 100 - 200$ (это те же 10 – 15 мс). Используем полученный минимальный сегмент данных $\mathbf{x}_\nu = \{x_1, ..., x_L\}$ в качестве обучающей выборки X_1 для оценивания автокорреляционной матрицы первой МЗЕ из сигнала. Соответствующий закон распределения $\mathbf{P}_1 = N(\hat{\mathbf{K}}_1)$ — это первый из элементов нашего будущего списка. После этого приравниваем $R = 1$ и берем второй сегмент выборки для анализа: $\mathbf{x}_2 = \{x_{L+1}, ..., x_{2L}\}$. Следуя выражению для решающей статистики МИР, определим для него удельную величину информационного рассогласования (ВИР) [2]

$$\rho(X_2, X_r) = \rho_r(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_2}, \; r = 1 \qquad (1)$$

относительно первой МЗЕ. Полученный результат сопоставляется с порогом по ВИР в роли допустимой величины рассогласований между разными реализациями одних и тех же фонем устной речи:

$$\rho(X_2, X_r) \leq \rho_0. \qquad (2)$$

Здесь ρ_0 — пороговый уровень. При нарушении данного неравенства в нашем начальном списке фонем появится второй элемент, и вслед за этим приравниваем число выявленных фонем $R = 2$. В противном случае принимается решение об объединении выборок X_1 и X_2 в один речевой образ $\mathbf{P}_1$ в качестве или одной МЗЕ удвоенной длительности $L_r = 2L$, если выборки смежные, или двух разных реализаций первой фонемы, если выборки не стыкуются. Равенство $R = 1$ в обоих случаях сохраняется.

Методика формирования ФБД

Вычисления по схеме (1), (2) повторяются циклически для всех последующих сегментов данных из речевого сигнала $X(t)$, причем повторяются «нарастающим итогом» для переменного значения $R = 2, 3, ...$. Каждый очередной сегмент данных сопоставляется по правилу (2) одновременно со всеми R множествами $\{X_r\}$ из текущего списка фонем. При этом не исключается возможность объединения одного и того же сегмента данных с элементами одновременно нескольких разных множеств. В результате будем иметь список фонем с некоторым фиксированным числом элементов R^*. Это важная характеристика как анализируемого речевого сигнала, так и самого диктора. Чем больше значение R^* для конкретного диктора, тем богаче с фундаментальной, фонетической точки зрения его речь. В данном выводе и состоит, как нам кажется, главный смысл и назначение фонетического анализа речи. Однако здесь же возникает и очевидная проблема: чрезмерно большое число фонем в речи диктора — это признак ее нечеткости или неинформативности. С точки зрения качества устной речи первостепенный интерес, безусловно, представляет собой множество четких МЗЕ. Его, в таком случае, и следует считать основным итогом фонетического анализа речи. Поэтому логика подсказывает: после выполнения всех перечисленных выше вычислений некоторые «фонемы» из окончательного списка можно исключить как маргинальные.

Добавим к сказанному, что рассматриваемая методика имеет множество разнообразных реализаций за счет, главным образом, применения рекуррентных вычислительных процедур корреляционно-спектрального анализа. Среди них наибольший интерес представляет метод обеляющего фильтра, основанный на авторегрессионной (АР) модели МЗЕ. В работах [1, 2] было показано, что в асимптотике, когда $n \to \infty$, и при гауссовом распределении речевого сигнала $\mathbf{P}_r = N(\mathbf{K}_r)$ с обратной автокорреляционной матрицей ленточной структуры выражение для оптимальной решающей статистики из выражения (1) сводится к виду

$$\rho_r(\mathbf{x}) = \frac{1}{F+1}\sum_{f=0}^{F}\frac{|A_r(jf)|^2}{|A_x(jf)|^2} - 1 \geq 0, \qquad (3)$$

где

$$A_r(jf) = 1 + \sum_{m=1}^{p} a_r(m) e^{-j\pi m f / F};$$

$$A_x(jf) = 1 + \sum_{m=1}^{p} a_x(m) e^{-j\pi m f / F}.$$

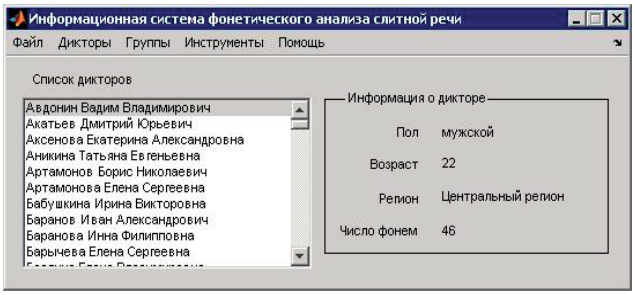
Здесь $\{a_x(m)\}$, $\{a_r(m)\}$ — два вектора АР-коэффициентов: входного сигнала и r-го эталона, оба одного порядка $p > 1$, а в числителе и знаменателе подынтегрального выражения отображены обратные зависимости спектральной плотности мощности (СПМ) соответственно для r-й фонемы, или ЭРЕ, и МЗЕ на входе. Это стандартная формулировка метода обеляющего фильтра в частотной области. Преимуществом данной интерпретации критерия МИР является, прежде всего, возможность его эффективной реализации в адаптивном варианте на основе быстрых вычислительных процедур АР-анализа, таких как метод Берга и др. Именно такой вариант метода обеляющего фильтра был реализован в дальнейшем для проведения его экспериментальных исследований в типовой задаче фонетического анализа речи.

Пример реализации

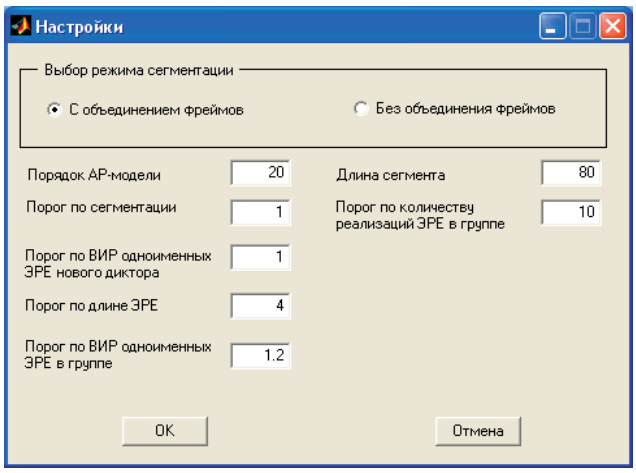
Для реализации предложенного алгоритма (1)—(3) была использована информационная система (ИС) фонетического анализа слитной речи [3]. На главном окне программы (рис. 1) отображаются главное меню и список дикторов, внесенных в БД. При выборе из списка любого диктора в правой части окна выводится краткая информация о нем.

Форма настроек ИС показана на рис. 2. Здесь задаются основные параметры для работы реализованных в ней алгоритмов.

Порядок АР-модели — целое число, большее единицы. Рекомендуется задавать его значение в пределах от 10 до 20. Порог по сегментации — это порог ρ_0 из выражения (2). Рекомендуется задавать в диапазоне от 0,7 до 1,5 (порог разладки при сегментировании должен быть больше 0,5). Этот порог используется на этапе сегментирования входного сигнала на фонемы. Порог по ВИР



■ ***Рис. 1.*** Главное окно программы ИС фонетического анализа слитной речи



■ ***Рис. 2.*** Форма настроек ИС

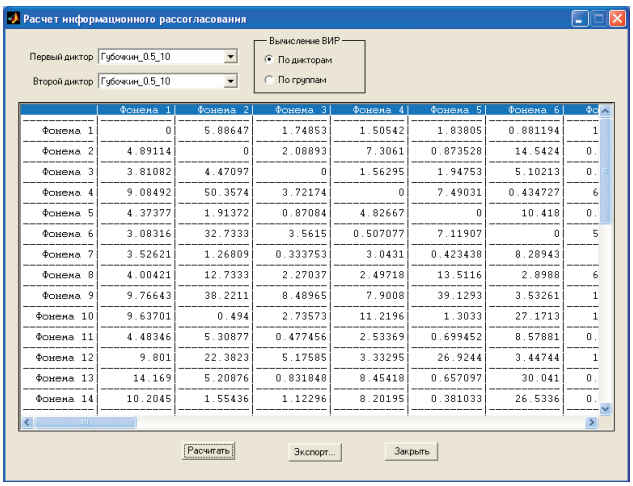
одноименных ЭРЕ нового диктора — любое число больше 0. Рекомендуется задавать в диапазоне от 0,8 до 2,0. Этот порог используется при объединении фонем, выделенных на этапе сегментирования в классы. Порог по длине ЭРЕ — целое число больше 0. Рекомендуется задавать в интервале от 3 до 7. Этот порог используется для задания минимального количества соседних МЗЕ или числа смежных сегментов, которые должна включать в себя фонема, для того чтобы она могла участвовать в процедуре классификации. Порог по ВИР одноименных ЭРЕ в группе — любое число больше 0. Рекомендуется задавать в интервале от 0,8 до 2,0. Этот порог используется при наполнении фонемами группы дикторов.

Длина сегмента задается в отсчетах и по умолчанию равна 80. Рекомендуется задавать ее в пределах от 80 до 320. Порог по количеству реализаций ЭРЕ в группе определяет минимальное количество выделенных реализаций, относящихся к одной фонеме, при котором данная фонема будет включена в БД. По умолчанию значение данного порога равно 10.

Форма расчета ВИР между фонемами разных дикторов или групп показана на рис. 3. Соответствующий режим выбирается кнопками «По дикторам» и «По группам».

Процесс создания ФБД на основе данной ИС выполняется в несколько этапов. На первом этапе формируется группа дикторов, и каждый из них проговаривает в среднем темпе лингвистически сбалансированный текст или отрывок из художественного произведения длительностью 1–2 мин. При этом объем текста составляет минимум 1–1,5 тыс. печатных знаков. Каждая такая запись с помощью звукового редактора сохраняется в виде соответствующего звукового файла.

На втором этапе экспериментальных исследований производится обработка полученных фай-



■ ***Рис. 3.*** Форма расчета ВИР между фонемами

лов по адаптивному алгоритму (1)—(3). В результате формируется множество персональных ФБД $\{X_r\}$, учитывающих особенности разных дикторов. Это главный результат автоматической обработки речевых сигналов.

На третьем, заключительном этапе обработки речевых сигналов отбирается для анализа несколько персональных ФБД. В пределах полученного множества осуществляется объединение отдельных элементов ФБД по принципу МИР общего вида (2). По результатам такого анализа делаются выводы об устойчивости объединенной ФБД к индивидуальным особенностям речи дикторов.

Основные результаты

Предложенная методика была реализована практически для группы дикторов [4], составленной из жителей севера Нижегородской области (всего 100 чел.) примерно одного возраста (25 – 30 лет) и одного пола (мужчины). Каждым диктором был проговорен тестовый текст объемом около одной стандартной машинописной страницы, взятый из первой главы романа А. С. Пушкина «Капитанская дочка». Частота дискретизации встроенного АЦП была установлена равной 8 кГц — общепринятое значение при обработке разговорной речи. Продолжительность записи по каждому диктору составила не менее 1,5 мин. При этом длина L одного сегмента данных во всех случаях устанавливалась равной 80 отсчетам, или 10 мс по времени, порядок АР-модели $p = 20$, а пороги по ВИР и длине ЭРЕ — $\rho_0 = 1{,}1$ и $L_0 = 320$ отсчетов соответственно. В результате обработки полученных записей согласно методике (1)—(3) было создано 100 персональных ФБД $\{\mathbf{x}_r^*\}$. После их объединения в одну ФБД результирующий список включил в себя $R_0 = 118$ фонем. Для подтверждения того, что фонемы полученной ФБД включают в себя фонемы всех дикторов из рассматриваемого множества, было проведено сопоставление объединенной и персональной ФБД одного из дикторов, которая содержала $R_1 = 57$ фонем. Сопоставление производилось по матрице (57×57) ВИР между однотипными (в смысле МИР) МЭЕ. Ее фрагмент представлен в табл. 1.

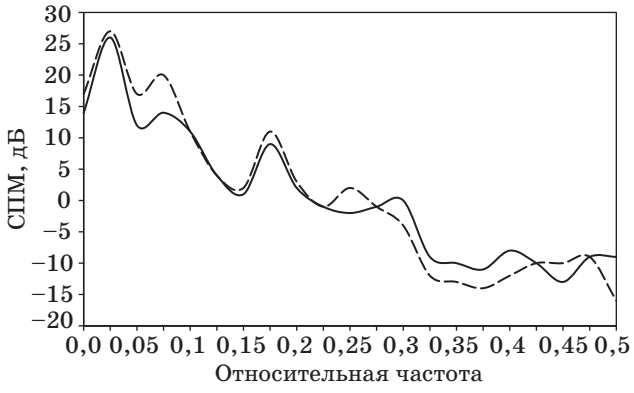
Из таблицы видно, что все диагональные элементы матрицы ВИР существенно меньше по величине, чем элементы, находящиеся вне ее главной диагонали. Это очевидный признак высокой степени подобия соответствующих фонем в теоретико-информационном смысле. Нулевые значения ВИР означают, что фонемы с данными номерами были включены в группу в качестве эталонных. В подтверждение этому на рис. 4 показаны графики СПМ первой фонемы выбранного нами для анализа диктора (сплошная линия) и наиболее близкой к ней по критерию МИР (1) фонемы из объединенного списка фонем (штриховая линия).

Видно, что обе СПМ практически не отличаются друг от друга. Отметим, что аналогичный результат достигается и для всех других пар одноименных фонем.

На заключительном этапе экспериментальных исследований рассматривались результаты формирования ФБД при более высоких значениях порогов по ВИР и длине ЭРЕ: $\rho_0 = 1{,}5$ и $L_0 = 400$. При этом вычисления проводились по той же схе-

■ ***Таблица 1***

Номер фонемы	1	2	3	.	53	54	55	56	57
1	0,382	1,905	3,637	.	1,263	4,583	3,515	2,066	10,67
2	1,290	0,333	0,564	.	9,815	20,09	4,483	10,17	115,5
3	3,928	1,113	0,381	.	18,32	40,74	6,846	18,35	274,3
.....									
55	2,796	1,778	1,568	·	9,703	6,763	0	6,109	59,72
56	1,463	2,523	3,079	·	4,202	27,39	5,834	0	65,64
57	10,62	10,05	19,05	·	51,84	67,90	25,45	17,41	0



■ ***Рис. 4.*** СПМ двух фонем

■ *Таблица 2*

Номер фонемы	1	2	3	.	23	24	25	26	27
1	0,47	1,22	2,78	·	4,50	15,08	14,85	7,19	4,31
2	0,50	0,26	2,42	·	1,16	13,50	18,60	4,54	3,54
3	0,94	0,89	0,43	·	3,20	4,29	32,46	3,88	10,67
.....									
25	6,99	10,10	5,57	.	30,10	42,31	0,910	16,43	12,54
26	2,72	2,06	1,21	·	16,54	8,76	12,71	0,09	6,02
27	1,42	1,59	2,06	·	4,11	18,30	6,19	5,45	0,54

ме (1)—(3). В результате был сформирован объединенный список $\{X_r\}$, содержащий $R_0 = 45$ фонем. После его сопоставления с ФБД (27×27) нашего первого диктора была получена матрица ВИР, фрагмент которой показан в табл. 2.

Из сопоставления табл. 1 и 2 можно сделать важный вывод о том, что вне зависимости от значений параметров настроек ИС объединенная ФБД, сформированная по предложенной методике, сохраняет в себе необходимую информацию об особенностях произношения каждого отдельного диктора из заданной группы.

Заключение

Известно, что в мире на данный момент не существует высококачественного программного продукта в области автоматического распознавания речи (АРР) на русском языке. Причина кроется в его исключительных лингвистических особенностях [5], а также в известных (см. ГОСТ Р 50840-95 и др.) жестких нормативных требованиях к системам передачи и обработки русской разговорной речи. До последнего времени данная проблема являлась главным препятствием на пути широкого распространения новых речевых технологий в России. И даже в самых передовых мировых разработках в области АРР, таких как Google Voice, Apple Siri и др., она до конца не преодолена: вероятность ошибки распознавания в них не опускается ниже 15–20 %. В отличие от существующих аналогов в предложенном выше исследовании была применена недавно созданная авторами информационная теория — совместно с кластерной моделью МЗЕ и общесистемным критерием МИР в информационной метрике Кульбака — Лейблера. На данный момент это весьма перспективное направление в рамках набирающего силу фонетического подхода [6–8] к задачам АРР на русском языке. Его основное преимущество перед известными подходами и методами АРР состоит в достигаемой полной автоматизации процедуры формирования ФБД, при этом резко сокращается и время на реализацию данного процесса, а это главное условие высокого быстродействия в целом системы АРР при работе в режиме реального времени.

Таким образом, отталкиваясь от ряда основных положений информационной теории восприятия речи, авторы предлагают новую методику формирования ФБД в автоматическом режиме, обладающую широкими функциональными возможностями и перспективами для практического применения.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ по государственному контракту № 07.514.11.4137 ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы».

Литература

1. **Савченко В. В.** Информационная теория восприятия речи // Изв. вузов. Радиоэлектроника. 2007. Вып. 6. С. 10–14.
2. **Савченко В. В.** Автоматическая обработка речи по критерию минимума информационного рассогласования на основе метода обеляющего фильтра // Радиотехника и электроника. 2005. Т. 50. № 3. С. 309–314.
3. **Свидетельство Роспатента РФ** о гос. рег. программы для ЭВМ. № 2008615442. Информационная система фонетического анализа слитной речи: программа для ЭВМ / В. В. Савченко, Д. Ю. Акатьев, И. В. Губочкин и др. Выдано 14.11.2008.
4. **Свидетельство Роспатента РФ** о гос. рег. базы данных. № 2009620512. Фонетическая база данных / В. В. Савченко, Д. Ю. Акатьев, И. В. Губочкин и др. Выдано 25.05.2009.
5. **Бабин Д. Н., Мазуренко И. Л., Холоденко А. Б.** Проблемы создания автоматического распознавателя слитной устной русской речи // Интеллектуальные системы в производстве. 2003. № 1. С. 4–23.
6. **Кодзасов С. В., Кривнова О. Ф.** Общая фонетика / РГГУ. — М., 2001. — 592 с.
7. **Ронжин А. Л., Ли И. В.** Автоматическое распознавание русской речи // Вестник Российской академии наук. 2007. Т. 77. № 2. С. 133–138.
8. **Кипяткова И. С., Карпов А. А.** Эксперименты по распознаванию слитной русской речи с использованием сверхбольшого словаря // Тр. СПИИРАН. 2010. Вып. 12. С. 63–74.