РАДИОТЕХНИКА И ЭЛЕКТРОНИКА, 2022, том 67, № 3, с. 286–293

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ СИГНАЛОВ

УДК 621.391:004.522

ОБНАРУЖЕНИЕ ГЛАСНЫХ ЗВУКОВ РЕЧИ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ С ГАРАНТИРОВАННОЙ НАДЕЖНОСТЬЮ

© 2022 г. А. В. Савченко[a], *, В. В. Савченко[b], **

[a] Национальный исследовательский университет "Высшая школа экономики", ул. Б. Печерская, 25, Нижний Новгород, 603155 Российская Федерация

[b] Редакция журнала "Радиотехника и электроника", ул. Моховая, 11, стр. 7, Москва, 125009 Российская Федерация

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

***E-mail: vvsavchenko@yandex.ru*

Поступила в редакцию 25.02.2021 г.
После доработки 21.10.2021 г.
Принята к публикации 28.10.2021 г.

Рассмотрена задача обнаружения гласных звуков речи в режиме реального времени. Предложен новый алгоритм для ее решения на основе информационного ($R + 1$)-элемента и метода обеляющего фильтра. Рассмотрен пример его практической реализации, даны оценки эффективности. Поставлен и проведен натурный эксперимент. Показано, что при минимальных требованиях к производительности используемой вычислительной техники предложенный алгоритм характеризуется достаточно высоким быстродействием и гарантированным уровнем значимости принимаемых решений.

DOI: 10.31857/S0033849422030135

ВВЕДЕНИЕ

Известно [1], что гласные звуки речи (ГЗР) представляют собой наиболее значимые речевые события как с точки зрения производства, так и с точки зрения анализа речи. Их обнаружение в составе непрерывного речевого сигнала относится к числу классических задач в области автоматической обработки речи (АОР) [2–4]. В последние годы эта задача привлекает повышенный интерес исследователей в связи с появлением и распространением в мире бимодальных информационных систем и технологий [5, 6]. В них ГЗР служат сигналами условному наблюдателю для концентрации его внимания на артикуляции пользователей в моменты вероятных перемен в их эмоциональном состоянии [7, 8]. Сейчас это одно из наиболее востребованных направлений исследований в области АОР [9, 10].

Ввиду известного эффекта вариативности речи диктора на фонетическом уровне ее восприятия [11, 12] задача традиционно формулируется в терминах проверки статистических гипотез [13]. Решение в ней в многоальтернативном варианте принимают по критерию максимума правдоподобия [14]. При этом допускают ошибки разного рода [15], а именно: 1) пропуск гласной фонемы, 2) ее ложное обнаружение и, наконец, 3) перепутывание двух гласных. Их вероятности зависят от множества факторов, включая фонетические особенности речи диктора и качество используемого канала связи. А "вес" или "стоимость" таких ошибок могут сильно разниться в зависимости от поставленной наблюдателем задачи. Так, например, при анализе эмоционального состояния пользователей в многомодальных информационных системах первостепенное значение имеют ошибки первого рода, поскольку перепутывание гласных фонем друг с другом, или даже с согласными, не связано в данном случае с риском серьезных потерь полезной информации. Поэтому вероятность ошибки первого рода, или уровень значимости принимаемых решений, может служить показателем надежности используемого алгоритма обнаружения ГЗР.

Сложность состоит в том, что гласные звуки далеко не исчерпывают собой всего фонетического многообразия речи диктора. Так, например, русский язык наряду с шестью гласными ($R = 6$) насчитывает около сорока других фонем и несколько сотен их аллофонов [16]. В задаче обнаружения ГЗР их следует рассматривать в качестве интенсивных акустических помех речеподобного типа [17], которые сильно осложняют ее решение в режиме реального времени. По-видимому, именно этим обстоятельством можно объяснить тот общеизвестный факт [15, 16], что до настоящего времени в мире не создан сколько-нибудь эффективный коммерческий образец обнаружите-

ля ГЗР. Поэтому актуальность темы проведенного далее исследования представляется очевидной.

Для решения похожей задачи в работе [17] было предложено расширить множество гипотез до $R_\Sigma > R$ единиц за счет принятия к рассмотрению $R_\Sigma - R$ дополнительных альтернатив, учитывающих множество речеподобных помех. Правда, в нашем случае такой вариант наталкивается на проблему множественных сравнений [14], когда недопустимо (по степенной зависимости от R_Σ) возрастает вероятность ошибки "ложной тревоги". Однако для решения данной проблемы в теории АОР разработан эффективный математический аппарат, а именно: информационный $(R+1)$-элемент [17, 18]. Это условный термин, введенный в работе [19] для обозначения устройства или алгоритма проверки статистических гипотез в пределах неполного множества (объема $R < R_\Sigma$) альтернативных распределений вероятности. В отличие от известных алгоритмов с R выходами информационный $(R+1)$-элемент имеет дополнительный, $(R+1)$-й выход, который используется наблюдателем для регистрации отказа одновременно от всех R контролируемых альтернатив. Указанная особенность открывает широкие возможности для преодоления проблемы множественных сравнений в задаче обнаружения ГЗР. Исследованию данных возможностей и их воплощению в алгоритм гарантированной надежности для его применения в режиме реального времени и посвящена настоящая статья. При этом используется методология информационной теории восприятия речи [20–22].

1. ПОСТАНОВКА ЗАДАЧИ

Отталкиваясь от распространенной в задачах АОР [22, 23] многомерной (n-мерной) гауссовой аппроксимации Norm($\mathbf{K}$) N-вектора отсчетов $\mathbf{x}$ (фрейма) речевого сигнала $x(t)$ на интервалах его приблизительной (квази) стационарности, рассмотрим задачу проверки двух статистических гипотез

$$\left.\begin{array}{ll} H: & \mathbf{K} \subset \{\mathbf{K}_r\} \\ \overline{H}: & \mathbf{K} \not\subset \{\mathbf{K}_r\} \end{array}\right\}$$

в отношении его закона распределения с автокорреляционной матрицей (АКМ) $\mathbf{K}$. Здесь $\mathbf{K}_r$ – АКМ r-й гласной фонемы (чертой над символом H обозначено логическое отрицание). Как видим, обе гипотезы являются сложными [13]. Задача в данной формулировке не имеет оптимального решения [14]. Проблема может быть преодолена путем сведения рассматриваемой задачи к R-кратной дихотомии [18]

$$\left.\begin{array}{ll} H_r: & \mathbf{K} = \mathbf{K}_r \\ \overline{H}_r: & \mathbf{K} \neq \mathbf{K}_r \end{array}\right\}, \quad r = \overline{1, R}, \qquad (1)$$

по числу гласных фонем в речи контрольного диктора. При этом гипотеза H принимается при условии справедливости любой из парциальных гипотез H_r, т.е. выполняется равенство

$$H = \bigcup_{r=1}^{R} H_r.$$

Так формулируется задача об обнаружении "разладки" в случайном гауссовом процессе [22]. В ней сложной остается только вторая (альтернативная) гипотеза. Теория рекомендует применять в подобных случаях критерии несмещенного типа, для которых вероятность ошибки первого рода не превышает вероятности ошибки второго рода. В задаче (1) в этом качестве можно использовать критерий отношения правдоподобия [13]

$$W_r(\mathbf{x}):\ \lambda_r(\mathbf{x}) \triangleq \frac{\sup_{\mathbf{K}_r} p_r(\mathbf{x})}{p_r(\mathbf{x})} \le \lambda_0, \quad r = \overline{1, R}, \qquad (2)$$

где $p_r(\mathrm{x})$ – функция правдоподобия гипотезы H_r (символом Δ над знаком равенства здесь обозначено равенство по определению).

Решение $\overline{W}_r(\mathbf{x})$ не в пользу данной гипотезы принимается в (2) при условии превышения порогового уровня $\lambda_0 > 1$ отношением двух функций правдоподобия: эмпирического распределения и его гипотетической (r-й) альтернативы Norm($\mathbf{K}_r$), сформированной по результатам предварительного корреляционного анализа сигнала-эталона одноименного звука речи [11, 20]. Величина порога λ_0 устанавливается наблюдателем исходя из равенства вероятности ошибки первого рода [21]

$$\alpha_r \triangleq P\left\{\overline{W}_r(\mathbf{x}) \middle| H_r\right\} = P\left\{\lambda_r(\mathbf{x}) > \lambda_0 \middle| H_r\right\} = \alpha_0, \qquad (3)$$

заданной константе $\alpha_0 \ll 1$, где $P\{\cdot|\cdot\}$ – условная вероятность случайного события. В таком случае правило (2) гарантирует требуемый уровень значимости принимаемых решений [13]. Причем не исключается возможность срабатывания данного критерия одновременно для нескольких гласных фонем с номерами $r_1, r_2, \ldots, r_L$, где $L \le R$. Однако суммарная вероятность ошибки первого рода при этом не увеличивается:

$$\alpha_\Sigma = \prod_{l=1}^{L} P\left\{\lambda_{r_L}(\mathbf{x}) > \lambda_0 \middle| H_r\right\} \le$$
$$\le P\left\{\lambda_r(\mathbf{x}) > \lambda_0 \middle| H_r\right\} \le \alpha_0,$$

поскольку по условиям задачи обнаружения (1) различение ГЗР между собой не предусмотрено. Решение об обнаружении гласной фонемы принимается в общем случае при условии $L \ge 1$. Нетрудно понять, что этим одновременно решаются проблемы как множественных сравнений, так и быстродействия обнаружителя. Однако реализации данного эффекта на практике препятствует

проблема нестабильности в широких пределах масштаба или амплитуды ГЗР на входе обнаружителя [22].

В самом деле, учитывая тот факт, что используемые в (1) эталоны хранятся в базе данных обнаружителя ГЗР в виде R-множества АКМ фонетических образцов $x_r(t)$ фиксированной амплитуды, нетрудно представить себе остроту указанной проблемы для практики АОР: при любой константе $c_r > 0$ в роли масштабного множителя должна выполняться система равенств

$$W_r(\mathbf{x}) = W_r(c_r\mathbf{x}), \quad r = \overline{1, R}. \tag{4}$$

В противном случае решающее правило (2) утрачивает свою работоспособность, поскольку вне зависимости от уровня значимости α_0 будем иметь согласно (3) парадоксальное требование к пороговому уровню обнаружителя: $\rho_0 \to \infty$. Для устранения этого препятствия модифицируем критерий (2), наделив его свойством масштабной инвариантности (4).

2. СИНТЕЗ АЛГОРИТМА

Основываясь на блочно-последовательной структуре наблюдаемого фрейма $\mathbf{x} = \{\mathbf{x}_m\}$ центрированного речевого сигнала, запишем выражение [20, 22]

$$\begin{aligned} p_r(\mathbf{x}) &= p_r(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M) = \prod_{m=1}^{M} p_r(\mathbf{x}_m) = \\ &= \left[(2\pi)^n |\mathbf{K}_r|\right]^{-0.5M} \exp\left(-0.5\sum_{m=1}^{M} \mathbf{x}_m^T \mathbf{K}_r^{-1} \mathbf{x}_m\right) = \\ &= \left[(2\pi)^n |\mathbf{K}_r|\right]^{-0.5M} \exp\left[-0.5M \operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\right], \end{aligned}$$

или в более компактном виде

$$\begin{aligned} &\ln p_r(\mathbf{x}) = \\ &= -0.5M^{-1}\left[\operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) + \ln|\mathbf{K}_r| + \ln(2\pi)n\right], \end{aligned} \tag{5}$$

где $\mathbf{S} \triangleq M^{-1}\sum_{m=1}^{M} \mathbf{x}_m\mathbf{x}_m^T$ – эмпирическая оценка АКМ речевого сигнала по M-выборке векторных наблюдений; $\mathbf{x}_m$ – n-вектор (столбец) отсчетов речевого сигнала $x(t)$ в пределах его m-го ($m \le M$) отрезка длительностью $\tau_0 = \tau/M$; $M = [N/n]$, (символами $\operatorname{tr}(\cdot)$ и $|\cdot|$ здесь обозначены соответственно след и определитель квадратной $(n \times n)$-матрицы, $[\cdot]$ – целая часть рационального числа, ^{T} – знак транспонирования). Например, при $\tau = 30$ мс, $F = 8$ кГц и $n = 20$ (типичные значения параметров для систем АОР [21, 22]) будем иметь $N = 30 \times 8 = 240$ и, следовательно, $M = 240/20 = 12$ непересекающихся отрезков сигнала $x(t)$.

Следуя принципу максимума правдоподобия [13], в предположении о неособенности и положительной определенности матрицы $\mathbf{S}$ из (5) будем иметь [22]

$$\begin{aligned} &\sup_{\mathbf{K}_r} \ln p_r(\mathbf{x}) = \\ &= -0.5M^{-1}\left[\operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) + \ln|\mathbf{K}_r| + n\ln(2\pi)\right]\Big|_{\mathbf{K}_r=\mathbf{S}} = \\ &= -0.5M^{-1}\left[n + \ln|\mathbf{S}| + n\ln(2\pi)\right] = \\ &= -0.5M^{-1}\left[\ln|\mathbf{S}| + n(\ln(2\pi) + 1)\right]. \end{aligned} \tag{6}$$

Выражения (2), (5) и (6) в совокупности приводят к равенству

$$\begin{aligned} \ln\lambda_r(\mathbf{x}) &= \ln\sup_{\mathbf{K}_r} p_r(\mathbf{x}) - \ln p_r(\mathbf{x}) = \\ &= 0.5M^{-1}\left\{\left[\operatorname{tr}\left(\mathbf{S}\cdot\mathbf{K}_r^{-1}\right) + \ln|\mathbf{K}_r| + n\ln(2\pi)\right] - \right. \\ &\quad \left. - \left[\ln|\mathbf{S}| + n(\ln(2\pi) + 1)\right]\right\} = \\ &= 0.5M^{-1}\left[\operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) - \ln\left|\mathbf{S}\mathbf{K}_r^{-1}\right| - n\right]. \end{aligned}$$

При его учете критерий (2) может быть переписан в эквивалентном виде

$$W_r(\mathbf{x}): \ \rho_r(\mathbf{x}) \le \rho_0, \tag{7}$$

где в качестве решающей статистики используется удельная величина (на один отсчет данных) информационного рассогласования

$$\rho_r(\mathbf{x}) \triangleq 0.5\left[n^{-1}\operatorname{tr}(\mathbf{S}\mathbf{K}_r^{-1}) - n^{-1}\ln\left|\mathbf{S}\mathbf{K}_r^{-1}\right| - 1\right] \tag{8}$$

двух гауссовых n-мерных распределений $\mathrm{Norm}(\mathbf{K}_r)$ и $\mathrm{Norm}(\mathbf{S})$ по Кульбаку–Лейблеру [24]. Ее пороговый уровень ρ_0 по аналогии с (3) определяется корнем уравнения

$$P\{\rho_r(\mathbf{x}) > \rho_0 | H_r\} = \alpha_0.$$

Выражения (7), (8) определяют в явном виде алгоритм обнаружения ГЗР в пределах наблюдаемого фрейма $\mathbf{x}$ речевого сигнала. Хотя он и не обладает свойством масштабной инвариантности в явном виде, математическая формулировка решающей статистики (8) открывает возможность для его достижения на основе использования апробированного в работе [25] подхода.

Следуя принципу минимума информационного рассогласования (МИР) [24, 26], рассмотрим оптимизационную задачу: найти минимум информационного рассогласования

$$\begin{aligned} \rho_r(c_r\mathbf{x}) &= 0.5\left[n^{-1}c_r^2\operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) - n^{-1}\ln\left|c_r^2\mathbf{S}\mathbf{K}_r^{-1}\right| - 1\right] = \\ &= 0.5\left[c_r^2 n^{-1}\operatorname{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) - \ln c_r^2 + \right. \\ &\quad \left. + n^{-1}\ln|\mathbf{S}^{-1}| - n^{-1}\ln\left|\mathbf{K}_r^{-1}\right| - 1\right] \end{aligned} \tag{9}$$

для нестабильного сигнала $\mathbf{x}$ по переменной величине его масштабного множителя $c_r > 0$. Для

этого сначала найдем для целевой функции задачи $\rho_r(c_r) \triangleq \rho_r(c_r\mathbf{x})\big|_{\mathbf{x}=\text{const}}$ первую производную:

$$\frac{\mathrm{d}\rho_r(c_r)}{\mathrm{d}c_r} = c_r n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) - c_r^{-1}.$$

Приравнивая ее нулю, получим оптимизационное уравнение

$$c_r^2 n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right) - 1 = 0,$$

решая которое, находим корень общего вида

$$c_r^* = \left[n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\right]^{-0.5}.$$

После подстановки полученного результата в выражение (9) будем иметь

$$\rho_r^*(\mathbf{x}) \triangleq \rho_r(c_r^*\mathbf{x}) = \\ = 0.5\left[\ln\left(n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\right) + n^{-1}\ln\left|\mathbf{S}^{-1}\right| - n^{-1}\ln\left|\mathbf{K}_r^{-1}\right|\right]. \quad (10)$$

Полученное выражение определяет решающую статистику МИР как альтернативу (8) для подстановки в критерий (7). Нетрудно увидеть, что эта статистика обладает свойством масштабной инвариантности в смысле равенства (4), а именно:

$$\forall c_r > 0: \quad \rho_r^*(c_r\mathbf{x}) = 0.5\left[\ln\left(n^{-1}\mathrm{tr}\left(c_r{}^2\mathbf{S}\mathbf{K}_r^{-1}\right)\right) + \right. \\ \left. + n^{-1}\ln\left|c_r^{-2}\mathbf{S}^{-1}\right| - n^{-1}\ln\left|\mathbf{K}_r^{-1}\right|\right] = \\ = 0.5\ln\left[c_r^2 n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\left|c_r^{-2}\mathbf{S}^{-1}\right|^{1/n}\left|\mathbf{K}_r^{-1}\right|^{-1/n}\right] = \\ = 0.5\ln\left[n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\left|\mathbf{S}^{-1}\right|^{1/n}\left|\mathbf{K}_r^{-1}\right|^{-1/n}\right] = \\ = 0.5\left[\ln\left(n^{-1}\mathrm{tr}\left(\mathbf{S}\mathbf{K}_r^{-1}\right)\right) + \right. \\ \left. + n^{-1}\ln\left|\mathbf{S}^{-1}\right| - n^{-1}\ln\left|\mathbf{K}_r^{-1}\right|\right] = \rho_r^*(\mathbf{x}).$$

Обоснованием алгоритма (7), (10) может служить и соображение практического характера: учитывая быструю сходимость (со скоростью неулучшаемого порядка $1/M \sim 1/N$) статистических оценок АКМ по формуле выборочного среднего [13, 24], можно ожидать, что эмпирическое распределение Norm(**S**) должно не сильно отличаться от своего эталона Norm($\mathbf{K}_r$) при справедливости гипотезы H_r в условиях конечных ($N < \infty$) выборок наблюдений. Раскроем принцип действия синтезированного алгоритма на примере его практической реализации с использованием распространенной в задачах АОР [20–23] авторегрессионной модели ГЗР.

3. ПРИМЕР ПРАКТИЧЕСКОЙ РЕАЛИЗАЦИИ

Авторегрессионная (АР) модель сигнала r-й фонемы

$$x_r(t) = \sum_{i=1}^{p} a_r(i)x_r(t-i) + \eta_r(t), \quad t = 1, 2, \ldots, \quad (11)$$

однозначно определяется своим вектором АР-коэффициентов $\mathbf{a}_r \triangleq \left\{a_r(i),\ i = \overline{1, p}\right\}$ конечного порядка p, а также дисперсией $\sigma_r^2 = \text{const}$ порождающего процесса $\{\eta_r(t)\}$ типа белого гауссова шума в дискретном времени t. С одной стороны, АР-модель (11) органично сочетается с голосовым механизмом человека (имеется в виду модель речевого тракта типа “акустическая труба” [1, 22]), с другой – существенно расширяет возможности программно-аппаратной реализации критерия (7). С указанной точки зрения представляют интерес известная [26–28] взаимосвязь АР-параметров речевого сигнала $x_r(t)$ и его АКМ $\mathbf{K}_r$.

Так, величина σ_r^2 определяет минимально достижимую дисперсию погрешности линейного предсказания случайного временного ряда (11) на один шаг в будущее. При условии $p < n$ она равна обратной величине первого элемента обратной одноименной (r-й) АКМ [27]:

$$\sigma_r^2 = \left(\mathbf{e}^T\mathbf{K}_r^{-1}\mathbf{e}\right)^{-1}.$$

Здесь символом **e** обозначен индикаторный вектор-столбец размерности n, составленный из одних нулей, за исключением единицы на первой позиции. Аналогичным образом может быть определен и соответствующий вектор АР-коэффициентов:

$$\left(1; -\mathbf{a}_r^T\right)^T = \sigma_r^2\mathbf{K}_r^{-1}\mathbf{e} = \frac{\mathbf{K}_r^{-1}\mathbf{e}}{\mathbf{e}^T\mathbf{K}_r^{-1}\mathbf{e}} \triangleq \mathbf{b}_r.$$

Он состоит из взятых с коэффициентом σ_r^2 элементов первого столбца обратной одноименной АКМ, исключая ее первый элемент. Здесь $\mathbf{b}_r$ – вектор коэффициентов линейного обеляющего фильтра (ОФ), настроенного на этапе подготовки данных на сигнал r-й фонемы $x_r(t)$. Его порядок равен $p = n - 1$. Данный фильтр – ключевой элемент обнаружителя ГЗР (7), (10).

Динамика ОФ описывается инверсным по отношению к (11) выражением вида [28]

$$y_r(t) = x(t) - \sum_{i=1}^{n-1} a_r(i)x(t-i), \quad t = 1, 2, \ldots \quad (12)$$

Дисперсия $\sigma_r^2(\mathbf{x}) \triangleq \left\langle y_r{}^2(t)\right\rangle$ сигнала на его выходе (скобками $\langle\cdot\rangle$ обозначено математическое ожидание случайной величины) отвечает соотношению $\sigma_r^2(\mathbf{x}) \geq \sigma_r^2$ [22] с равенством лишь в асимптотике (при $N \to \infty$), когда на вход r-го ОФ (12) поступает сигнал $x_r(t)$ одноименной фонемы. Эмпирическая (по выборке) оценка данной дисперсии определяется по формуле [13]

$$\hat{\sigma}_r^2(\mathbf{x}) = M^{-1}\sum_{m=1}^{M} y_r^2(\mathbf{x}_m) \quad (13)$$

выборочного среднего квадрата отклика $y_r(\mathbf{x}_m) = \mathbf{b}_r^T \mathbf{x}_m$ r-го ОФ на m-й отрезок $\boldsymbol{x}_m$ речевого сигнала (напомним, он предварительно центрирован). Дополним выражение (13) известным асимптотическим равенством [27]

$$n^{-1} \ln |\mathbf{K}_r| \Big|_{n\to\infty} = \ln \sigma_r^2,$$

а также его двумя статистическими аналогами [28]:

$$n^{-1} \mathrm{tr}(\mathbf{S}\mathbf{K}_r^{-1}) \Big|_{n\to\infty} = \frac{\hat{\sigma}_r^2(\mathbf{x})}{\sigma_r^2}, \quad n^{-1} \ln |\mathbf{S}| \Big|_{n\to\infty} = \ln \hat{\sigma}_x^2(\mathbf{x}),$$

где $\hat{\sigma}_x^2(\mathbf{x}) = (\mathbf{e}^T \mathbf{S}^{-1} \mathbf{e})^{-1}$ – эмпирическая дисперсия речевого сигнала на выходе адаптивного ОФ, настроенного по выборке наблюдений $\mathbf{x}$ в режиме "скользящего окна" длиной τ в один речевой фрейм. При их учете из выражения (10) будем иметь

$$\rho_r^*(\mathbf{x}) = 0.5 \ln \left[\frac{\sigma_r^2}{\hat{\sigma}_x^2(\mathbf{x})} \frac{\hat{\sigma}_r^2(\mathbf{x})}{\sigma_r^2} \right] = 0.5 \ln \left[\frac{\hat{\sigma}_r^2(\mathbf{x})}{\hat{\sigma}_x^2(\mathbf{x})} \right]. \quad (14)$$

Полученное выражение совместно с критерием (7) определяет искомый алгоритм обнаружения ГЗР на основе метода ОФ и принципа МИР со свойством масштабной инвариантности (4). Его вычислительная сложность имеет порядок n^3, что следует из известной [27] оценки затрат на операцию обращения симметричной $(n \times n)$-матрицы $\mathbf{S}$. Это совсем немного, если учесть, что в задачах АОР размерность распределения речевого сигнала ограничена величиной $n = 10...20$ [21–23].

4. АНАЛИЗ ЭФФЕКТИВНОСТИ

Оценим верхнюю границу решающей статистики (14):

$$\rho_r^*(\mathbf{x}) = 0.5 \ln \left[\frac{\hat{\sigma}_r^2(\mathbf{x})}{\hat{\sigma}_x^2(\mathbf{x})} \right] \le 0.5 \left[\frac{\hat{\sigma}_r^2(\mathbf{x})}{\hat{\sigma}_x^2(\mathbf{x})} - 1 \right] \triangleq \sup \rho_r^*(\mathbf{x}).$$

По этой границе из выражения (3) определим гарантированный уровень значимости принимаемых согласно (7) решений:

$$\alpha_r = P\left\{ \rho_r^*(\mathbf{x}) > \rho_0 \middle| H_r \right\} \le P\left\{ \sup \rho_r^*(\mathbf{x}) > \rho_0 \middle| H_r \right\} = P\left\{ \frac{\hat{\sigma}_r^2(\mathbf{x})}{\hat{\sigma}_x^2(\mathbf{x})} > 1 + 2\rho_0 \middle| H_r \right\}. \quad (15)$$

Учитывая, что обе эмпирические дисперсии в (15) рассчитываются по формуле среднего квадрата случайной гауссовой величины (13), по аналогии с работой [22] воспользуемся для их описания двумя χ^2-распределениями (Пирсона) с M-степенями свободы каждое. В предположении об их статистической независимости [28, 29] получаем

$$\alpha_r \le P\left\{ \frac{\chi_1^2(M)}{\chi_2^2(M)} > 1 + 2\rho_0 \middle| H_r \right\} = 1 - \Phi_{M,M}(1 + 2\rho_0), \quad (16)$$

где $\Phi_{M,M}(\cdot)$ – интегральная функция F-распределения Фишера с (M, M)-степенями свободы [13]. Значения последней подробно табулированы, в том числе в электронном виде. Особо отметим, что правая часть выражения (16) не зависит от номера фонемы r и поэтому распространяется на весь фонетический строй контрольного диктора. Приравнивая ее заданному уровню значимости α_0, получим выражение для требуемого порогового уровня

$$\rho_0 = 0.5\left[\Phi_{M,M}^{-1}(1 - \alpha_0) - 1 \right] \quad (17)$$

решающей статистики (14) для его подстановки в правую часть критерия (7).

Отметим, что этим действием условный наблюдатель устанавливает требования к уровню значимости принимаемых им решений [22]: чем меньше вероятность α_0, тем ниже требования наблюдателя к обнаружителю ГЗР и, следовательно, ниже значимость или надежность его решений. И, наоборот, при понижении порога ρ_0 уровень значимости возрастает. Например, при равенствах $\alpha_0 = 0.05$ и $M = 12$ с использованием электронных таблиц Excel будем иметь $\rho_0 = 0.84$. Отметим, что межфонемная величина информационного рассогласования (8) звуков речи диктора этот порог превышает на порядок и более [11, 29]. Таким образом, предложенный алгоритм может быть охарактеризован гарантированной надежностью обнаружения ГЗР в смысле уровня значимости принимаемых в нем решений. Для сравнения: его известные аналоги [30, 31], основанные на методе ОФ в формулировке (7), (8), подобным качеством не обладают, поскольку их решающие статистики свойством масштабной инвариантности не наделены.

Действительно, в этом случае из (10) будем иметь

$$\rho_r(\mathbf{x}) = 0.5\left[\frac{\hat{\sigma}_r^2(\mathbf{x})}{\sigma_r^2} + \ln \sigma_r^2 + c \right],$$

где $c = \mathrm{const}$, или в упрощенном виде [26]

$$\rho_r(\mathbf{x}) = 0.5\left[\hat{\sigma}_r^2(\mathbf{x}) + c \right]$$

– при дополнительной нормировке дисперсии σ_r^2 порождающего процесса в рамках АР-модели (11) к единичному уровню. Основываясь на той же, что и при выводе выражений (16), (17), χ^2-аппроксимации нормированной случайной величины

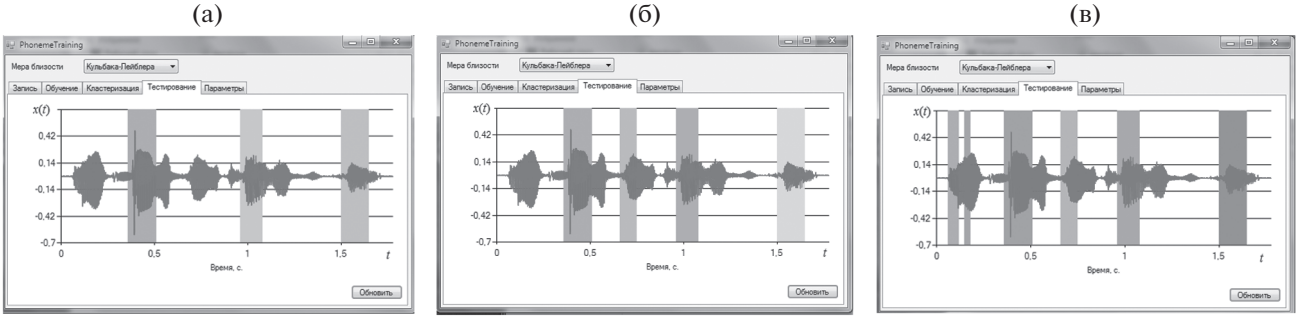

Рис. 1. Экранная форма рабочего окна программы “Сегментирование” при $\rho_0 = 0.60$ (а), 0.70 (б) и 0.85 (в).

$z^2(\mathbf{x}) \triangleq M\,\hat{\sigma}_r^2(\mathbf{x})/\sigma_r^2 = \chi^2(M)$, для этого случая можно записать

$$\alpha_r = P\{\rho_r(\mathbf{x}) > \rho_0 | H_r\} =$$
$$= P\left\{0.5\left[\hat{\sigma}_r^2(\mathbf{x}) + c\right] > \rho_0 \middle| H_r\right\} =$$
$$= P\left\{\hat{\sigma}_r^2(\mathbf{x}) > 2\tilde{\rho}_0 \middle| H_r\right\} =$$
$$= P\left\{\frac{\hat{\sigma}_v^2(\mathbf{x})}{\sigma_r^2(\mathbf{x})} > 2M\,\frac{\tilde{\rho}_0}{\sigma_r^2(\mathbf{x})}\middle| H_r\right\} =$$
$$= P\left\{\chi^2(M) > 2M\,\frac{\tilde{\rho}_0}{\sigma_r^2(\mathbf{x})}\right\} = 1 - \Phi_M\left(2M\,\frac{\tilde{\rho}_0}{\sigma_r^2(\mathbf{x})}\right).$$

Здесь $\Phi_M(\cdot)$ – интегральная функция χ^2-распределения с M-степенями свободы; $\tilde{\rho}_0 = \rho_0 + \text{const}$. Отсюда по аналогии с (17) получим выражение для требуемого порога

$$\tilde{\rho}_0 = \frac{\sigma_r^2(\mathbf{x})}{2M}\Phi_M^{-1}(1-\alpha_0)$$

в зависимости от установленного наблюдателем уровня значимости α_0. Как видим, в отличие от выражения (17), этот порог зависит не только от номера фонемы r, но и от интенсивности наблюдаемого речевого фрейма $\mathbf{x}$. Иными словами, при применении метода ОФ в формулировке (7), (8) нельзя гарантировать высокую степень надежности принимаемых наблюдателем решений. Напротив, предложенный алгоритм (7), (14) предоставляет наблюдателю такую возможность – путем регулировки уровня значимости в широком диапазоне значений α_0. Проиллюстрируем данную возможность результатами проведенного далее эксперимента.

5. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

Объектом проведенного исследования служил речевой сигнал $x(t)$ достаточно большой суммарной длительности (минуты), который был получен от контрольного диктора по результатам его устного чтения текста первой главы повести А.С. Пушкина “Капитанская дочка”. Предмет исследования – обнаружитель ГЗР, заданный своим критерием (7) и выражением для решающей статистики (14). Его реализация в программном виде была осуществлена на базе авторской компьютерной программы “Phoneme Training”[1]. Ее интерфейс ранее был подробно описан [20, 22]. Речевой сигнал $x(t)$ в ходе эксперимента членился на фреймы длительностью $\tau = 30$ мс – с 10 мс пересечениями каждого из них со своими “соседями” слева и справа. Частота дискретизации сигнала была установлена равной $F = 8$ кГц при равенстве порядка АКМ $n = 20$.

На этапе подготовки эксперимента по известной методике [29] была сформирована фонетическая база данных контрольного диктора. В нее вошли образцы (основные аллофоны [17]) его шести гласных фонем ($R = 6$). Длительность каждого образца составляла минимум $T = 2...3$ с. По ним сначала были получены оценки АКМ $\mathbf{K}_r$ для шести гласных фонем и сразу после этого – векторы коэффициентов $\mathbf{b}_r$ для системы ОФ (12). При этом точность полученных оценок в ее относительном выражении $\varepsilon = 1.65/\sqrt{3T/\tau}$ [21, 22] с доверительной вероятностью 0.9 не вышла за пределы 10%. Программа далее была переведена в режим “Сегментирование”, в котором речевой сигнал $x(t)$ был обработан согласно критерию (7), (14) с автоматическим выделением отрезков гласных фонем. Значение порогового уровня ρ_0 варьировалось в эксперименте с использованием вкладки “Параметры” в меню программы. Полученные результаты отражены на рис. 1 в виде трех скриншотов (снимков экрана) с изображением рабочего окна программы для разных значений порога ρ_0.

Здесь в каждом скриншоте представлена временная диаграмма сигнала короткого фрагмента речи диктора: “*Нас было девять человек детей*”. Серым цветом отмечены те отрезки речевого сиг-

[1] В настоящее время программа в режиме открытого доступа размещена на сайте https://sites.google.com/site/frompldcreators/produkty-1/phonemetraining.

нала, которые были идентифицированы обнаружителем как гласные фонемы. Разные оттенки серого отвечают разной степени надежности обнаружения ГЗР (зависит от мгновенного значения (14) решающей статистики МИР). Из сравнения временных диаграмм между собой можно подтвердить сделанный ранее вывод в отношении обратно пропорциональной зависимости (16) уровня значимости решений обнаружителя от установленного в нем порога ρ_0. Сделанный вывод имеет очевидное практическое значение с точки зрения оперативной регулировки порогового уровня условным наблюдателем под фонетические особенности диктора и национального языка.

6. ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Говоря о перспективах практического применения предложенного алгоритма, следует подробнее остановиться на задаче анализа динамики эмоционального состояния пользователей бимодальных (аудио и видео) информационных систем [32, 33]. Пусть нами выбран определенный алгоритм принятия решений, например, по артикуляции диктора в процессе речеобразования. Охарактеризуем его эффективность вероятностью безошибочного решения $P(A)$. Соответственно, $P(\bar{A}) = 1 - P(A)$ – вероятность ошибочного решения. Обозначим через $P(E)$ вероятность появления (на интервале наблюдений) ГЗР. Тогда $P(\bar{E}) = 1 - P(E)$ – это вероятность их отсутствия. По формуле полной вероятности [13] будем иметь

$$P(A) = P(AE) + P\left(A\overline{E}\right) =$$
$$= P(E)\times P(A|E) + P\left(\overline{E}\right)\times P\left(A|\overline{E}\right).$$

Из математической лингвистики известно [15, 16], что вероятности $P(E)$ и $P(\bar{E})$ сопоставимы между собой по величине, если не считать пауз в речи диктора между словами, а $P(A|E) \gg P(A|\bar{E})$, поскольку наиболее ярко эмоции проявляются через гласные [34, 35]. Поэтому в первом приближении можно записать $P(A) \simeq P(E)\times P(A|E)$, или $P(A|E) \simeq P(A)/P(E) > P(A)$.

Таким образом, при применении бимодального метода с обнаружением гласных фонем как стимула для концентрации внимания наблюдателя мы получаем гарантированный выигрыш по вероятности безошибочных решений. Величина выигрыша зависит от отношения двух условных вероятностей $\mu \triangleq P(A|E)/P\left(A|\overline{E}\right)$. При учете соотношения $P(A|E) \gg P(A|\bar{E})$ получаем выигрыш $\mu \gg 1$. Как видим, он может быть весьма значительным. И это только подтверждает тот общеизвестный факт [36, 37], что с точки зрения проявления эмоций в артикуляции диктора гласные звуки заведомо более информативны по сравнению со всеми другими звуками его речи.

К сожалению, в мире на данный момент не существует [38] коммерческого образца информационной системы, где этот эффект реализован на практике. Проблема состоит в организации работы обнаружителя ГЗР в режиме реального времени. Она обусловлена большой вычислительной сложностью существующих алгоритмов [2–6]. На решение этой проблемы и нацелен, главным образом, предложенный в рамках настоящей статьи алгоритм.

ЗАКЛЮЧЕНИЕ

Таким образом, благодаря проведенному исследованию предложен новый алгоритм обнаружения ГЗР для применения в режиме реального времени с регулируемым уровнем значимости принимаемых решений.

ФИНАНСИРОВАНИЕ

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 20-71-10010).

СПИСОК ЛИТЕРАТУРЫ

1. *Rabiner L.R., Shafer R.W.* Theory and Applications of Digital Speech Processing. Boston: Pearson, 2010.
2. *Kashani H.B., Sayadiyan A., Sheikhzadeh H.* // Speech Communication. 2017. V. 91. P. 28. https://doi.org/10.1016/j.specom.2017.04.008
3. *Srinivas N., Pradhan G., Kumar P.K.* // Integration. 2018. V. 63. P. 185. https://doi.org/10.1016/j.vlsi.2018.07.005
4. *Kumar A., Shahnawazuddin S., Pradhan G.* // Int. Conf. on Signal Processing and Communications (SPCOM). Bangalore. 16–19 Jul. 2018. N.Y.: IEEE, P. 252. https://doi.org/10.1109/SPCOM.2018.8724428
5. *Yongda D., Fang L., Huang X.* // Computers & Electrical Engineering. 2018. V. 72. P. 443. https://doi.org/10.1016/j.compeleceng.2018.09.014
6. *Hossain M.Sh., Muhammad G.* // Inform. Fusion. 2019. V. 49. P. 69. https://doi.org/10.1016/j.inffus.2018.09.008
7. *Akçay M.B., Oğuz K.* // Speech Communication. 2020. V. 116. P. 56. https://doi.org/10.1016/j.specom.2019.12.001
8. *Makino R., Yoshitomi Y., Asada T., Tabuse M.* // Proc. Int. Conf. on Artificial Life and Robotics. (ICAROB 2020). Oita. 8–11 Jan. Oita: Sugisaka Masanori, 2020. P. 403. https://doi.org/10.5954/ICAROB.2020.OS16-4
9. *Asada T. Adachi R., Takada S. et al.* // Proc. Int. Conf. on Artificial Life and Robotics. (ICAROB 2020). Oita. 8–11 Jan. Oita: Sugisaka Masanori, 2020. P. 398. https://doi.org/10.5954/ICAROB.2020.OS16-3

10. *Kumar A., Shahnawazuddin S., Pradhanet G.* // Circuits Systems, Signal Process. 2017. V. 36. P. 2315. https://doi.org/10.1007/s00034-016-0409-1
11. *Savchenko V.V.* // Radioelectron. Commun. Syst. 2020. V. 63. P. 532. https://doi.org/10.3103/S0735272720100039
12. *Lehet M., Holt L.* // Cognition. 2020. V. 202. P. 104328. https://doi.org/10.1016/j.cognition.2020.104328
13. *Боровков А.А.* Математическая статистика [Электронный ресурс]. Санкт-Петербург: Лань, 2010. https://e.lanbook.com/book/3810.
14. *Lehmann E.L., Romano J.P.* Testing Statistical Hypotheses. N. Y.: Springer, 2005. P. 348. https://doi.org/10.1007/0-387-27605-X
15. *Kashani H.B., Sayadiyan A.* // Computer Speech and Language. 2018. V. 50. P. 105. https://doi.org/10.1016/j.csl.2017.12.008
16. *Gahl S., Baayen R.H.* // J. Phonetics. 2019. V. 74. P. 42. https://doi.org/10.1016/j.wocn.2019.02.001
17. *Савченко В.В.* // РЭ. 2016. № 12. С. 1196. https://doi.org/10.7868/S0033849416120238
18. *Савченко В.В.* // Электросвязь. 2017. № 12. С. 22.
19. *Савченко В.В.* // Изв. вузов. Радиоэлектроника. 2006. № 4. С. 13.
20. *Савченко В.В.* // РЭ. 2019. Т. 64. № 6. С. 585. https://doi.org/10.1134/S0033849419060093
21. *Савченко В.В., Савченко Л.В.* // Измерительная техника. 2019. № 9. С. 59. https://doi.org/10.32446/0368-1025it.2019-9-59-64
22. *Савченко В.В., Савченко А.В.* // РЭ. 2020. Т. 65. № 11. С. 1101. https://doi.org/10.31857/S0033849420110157
23. *Candan Ç.* // Signal Processing. 2020. V. 166. P. 107256. https://doi.org/10.1016/j.sigpro.2019.107256
24. *Kullback S.* Information Theory and Statistics. N.Y.: Dover Publications, 1997. https://www.amazon.com//dp/0486696847.
25. *Savchenko A.V., Savchenko V.V. & Savchenko L.V.* // Optimization Lett. 2021. № 7. https://doi.org/10.1007/s11590-021-01790-5
26. *Савченко В.В.* // РЭ. 2005. Т. 50. № 3. С. 309.
27. *Marple S.L.* Digital Spectral Analysis with Applications. Mineola: Dover Publications. 2019. https://www.goodreads.com/book/show/19484239.
28. *Савченко В.В.* // РЭ. 1997. Т. 42. № 4. С. 426.
29. *Savchenko V.V.* // Radioelectronics and Communications. 2018. P. 61. № 9. P. 419. https://doi.org/10.3103/S0735272718090042
30. *Larsen B.S., Winther S., Nissen L. et al.* // Computing in Cardiology (CinC). Singapore, 8–11 Sept. 2019. N.Y.: IEEE, 2019. P. 9005907. https://doi.org/10.23919/CinC49843.2019.9005907
31. *Леховицкий Д.И., Атаманский Д.В., Рачков Д.С., Семеняка А.В.* // Изв. вузов. Радиоэлектроника. 2015. Т. 58. № 12(642). С. 3. https://doi.org/10.20535/S0021347015120018
32. *Akbulut F.P., Perros H.G., Shahzad M.* // Computer Methods and Programs in Biomedicine. 2020. V. 195. P. 105571. https://doi.org/10.1016/j.cmpb.2020.105571
33. *Falagiarda F., Collignon O.* // Cortex. 2019. V. 119. P. 184. https://doi.org/10.1016/j.cortex.2019.04.017
34. *Davis S.K., Morningstar M., Dirks M.A., Qualter P.* // Personality and Individual Differences. 2020. V. 160. P. 109938. https://doi.org/10.1016/j.paid.2020.109938
35. *Arana J., Gordillo F., Darias J., Mestas L.* // Computers in Human Behavior. 2020. V. 104. P.106156. https://doi.org/10.1016/j.chb.2019.106156
36. *Stasak B., Epps J., Goecke R.* // Computer Speech and Language. 2019. V. 53. P. 140. https://doi.org/10.1016/j.csl.2018.08.001
37. *Kim J., Toutios A., Lee S., Narayanan Sh.S.* // Computer Speech and Language. 2020. V. 64. P. 101100. https://doi.org/10.1016/j.csl.2020.101100
38. *Rammohan R., Dhanabalsamy N., Dimov V., Eidelman F.J.* // J. Allergy and Clinical Immunology. 2017. V. 139. № 2. P. AB250. https://doi.org/10.1016/j.jaci.2016.12.804