РАДИОТЕХНИКА И ЭЛЕКТРОНИКА, 2019, том 64, № 6, с. 585–592

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ СИГНАЛОВ

УДК 621.391:004.934

ДИВЕРГЕНЦИЯ ИТАКУРЫ–САИТО КАК ЭЛЕМЕНТ ИНФОРМАЦИОННОЙ ТЕОРИИ ВОСПРИЯТИЯ РЕЧИ

© 2019 г. В. В. Савченко*

Нижегородский государственный лингвистический университет, Российская Федерация, 603155 Нижний Новгород, ул. Минина, 31а

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

Поступила в редакцию 22.01.2018 г.
После доработки 24.10.2018 г.
Принята к публикации 01.11.2018 г.

На основе информационной теории восприятия речи дано обоснование симметричной формы дивергенции Итакуры–Саито в роли минимальной решающей статистики асимптотически оптимального алгоритма распознавания речевых сигналов на базовом, фонетическом уровне их обработки. Выводы теоретического исследования подтверждены результатами проведенного эксперимента. Показано, что благодаря применению синтезированного алгоритма удается существенно повысить точность и надежность автоматического распознавания наиболее проблемных фонетических единиц.

DOI: 10.1134/S0033849419060093

ВВЕДЕНИЕ

На протяжении многих лет дивергенция Итакуры–Саито (ДИС) широко используется в роли решающей статистики в системах автоматической обработки и распознавания речи (АОРР) [1–3]. Интерес к ней особенно возрос в последние годы в связи с тем, что такие признанные мировые лидеры в области АОРР, как Google и Apple Inc., осуществили свои прорывные речевые разработки по технологии “клиент–сервер” [4]. Сказанное объясняется целым рядом замечательных свойств ДИС, а именно: ее высокой чувствительностью к искажениям в спектрах сигналов, избирательностью по частоте, помехоустойчивостью, быстродействием и другими [5]. Однако ключевым является [6–8] свойство ДИС сочетаться наилучшим образом со средней экспертной (аудиторской) оценкой MOS (mean opinion score) акустического качества речи диктора, которая зарегистрирована в Международном союзе электросвязи МСЭ-Т в роли стандарта P.1120[1].

Между тем до настоящего времени распространение ДИС на практике не имеет строгого теоретического обоснования, если не считать ее изначальную связь [8, 9] с критерием максимума правдоподобия в задачах проверки статистических гипотез. Поэтому представляет интерес предлагаемое далее исследование, в рамках которого ДИС впервые теоретически обосновывается условиями априорной неопределенности в задаче распознавания образов. При этом используется математический аппарат информационной теории восприятия речи (ИТВР) [10–12], в которой, отталкиваясь от гауссовской аппроксимации речевых сигналов и принципа минимума их информационного рассогласования (МИР) по Кульбаку–Лейблеру [13], была впервые эффективно решена проблема априорной неопределенности. В развитие ряда идей и общих положений ИТВР в данной статье на основе принципа МИР дается вывод и обоснование симметричной формы ДИС как элемента асимптотически оптимального алгоритма АОРР. Выводы теоретического исследования подтверждены и проиллюстрированы результатами проведенного эксперимента.

1. ПОСТАНОВКА ЗАДАЧИ

Работа современных систем речевой обработки сводится, как правило [1–5], к поэлементному (на фонетическом уровне) сопоставлению произнесенного диктором слова или фразы $\mathbf{x}$ (n-вектор отсчетов речевого сигнала) с заранее подготовленным набором $\{\mathbf{x}_r\}_1^R$ эталонов сигналов фонем из числа их наблюдаемых реализаций (аллофонов). Здесь $R > 1$ – объем фонетической базы данных диктора. Решение принимается в пользу той фонемы, которая представляется гипотетическому слушателю ближе других к произнесенному слову $\mathbf{x}$ в некоторой метрике $\rho(\mathbf{x}/\mathbf{x}_r)$. При этом традиционно [5–8] используются метрики веро-

[1] См. ссылку в Интернете https://www.itu.int/ru/ITU-T/Pages/default.aspx.

ятностного типа. Задача в общем случае формулируется в терминах проверки R статистических гипотез в отношении закона распределения $\mathbf{P}_x$ наблюдаемого сигнала $\mathbf{x}$. Решение в пользу гипотезы $H_\nu: \mathbf{P}_x = \mathbf{P}_\nu$, $\nu \le R$, в ней принимается по признаку минимума решающей статистики общего вида

$$\rho(\mathbf{x}/\mathbf{x}_\nu) = \min_{r \le R} \rho(\mathbf{x}/\mathbf{x}_r), \quad (1)$$

определенной на R-множестве фонетических образцов $\{\mathbf{x}_r\}$. Так, при условии центрированности (нулевое среднее значение) и гауссовской аппроксимации речевого сигнала на интервалах его квазистационарности конечной длительности $\tau = \text{const}$ [6, 7] закон распределения $\mathbf{P}_r = \text{Norm}(\mathbf{K}_r)$ однозначно определяется набором неособенных автокорреляционных матриц (АКМ) $\mathbf{K}_r, r = \overline{1, R}$, конечной размерности $n \times n$. В расчете на использование при обработке речевого сигнала моментов, как правило, не выше второго порядка в работах [10–12] гауссовский закон был строго обусловлен общесистемным принципом максимума энтропии. В таком случае набор оптимальных решающих статистик в правой части критерия (1) определяется неотрицательной величиной информационного рассогласования (ВИР) Кульбака–Лейблера [13]

$$\rho(\mathbf{x}/\mathbf{x}_r) = \\ = \frac{1}{2n}\left[\text{tr}\left(\mathbf{S}_x \cdot \mathbf{K}_r^{-1}\right) - \ln\left|\mathbf{S}_x \cdot \mathbf{K}_r^{-1}\right| - n\right] \triangleq \rho_{x,r} \ge 0, \quad (2)$$

которая равна нулю лишь при условии $\mathbf{S}_x = \mathbf{K}_r$, где $\mathbf{S}_x$ – выборочная оценка АКМ наблюдаемого сигнала (символом “дельта” над знаком равенства здесь обозначено равенство по определению). При заданном априори наборе образцов $\{G_r(f)\}$ спектральной плотности мощности (СПМ) путем предельного перехода (при $n \to \infty$) из выражения (2) в частотной области получим [14]

$$\rho_{x,r} = F^{-1} \times \\ \times \int_{-F/2}^{F/2} \left(G_x(f)/G_r(f) - \ln\left(G_x(f)/G_r(f)\right) - 1\right) df, \quad (3)$$
$$r = \overline{1, R},$$

где F – частота дискретизации сигнала во времени, а $G_x(f)$ – его оценка СПМ по выборке с использованием известного [15] математического аппарата. Это стандартная [2, 3] формулировка критерия МИР в терминах ИТВР. Одновременно выражение (3) определяет ДИС для СПМ двух сигналов [1–6].

Проблема состоит [10, 11] в вариативности устной речи диктора, причем, в пределах даже одного речевого потока. Применительно к гауссовской модели речевого сигнала $\text{Norm}(\mathbf{K}_r)$ указанная вариативность порождает острейшую [15, 16] в задачах АОРР проблему априорной неопределенности в отношении спектрально-корреляционных свойств речевого сигнала. Задача (1)–(3) в таком случае сводится к распознаванию речевых образов [17–19].

2. СИНТЕЗ ОПТИМАЛЬНОГО АЛГОРИТМА

Следуя классическому критерию отношения правдоподобия [17], рассмотрим R классифицированных (на множестве эталонов фонем $\{\mathbf{x}_r\}$) многомерных (размера n) повторных (объема M) независимых выборок $\mathbf{x}_{r,j} = \left(x_{r,j}(1), x_{r,j}(2), \ldots, x_{r,j}(n)\right)^{\mathrm{T}}$ из R гауссовских популяций $\mathbf{P}_r = \text{Norm}(\mathbf{K}_r)$ с нулевыми математическими ожиданиями и неизвестными в общем случае АКМ $\mathbf{K}_r$, $r = \overline{1, R}$. Обозначим их в совокупности через $(n \times M)$-матрицу наблюдений $\mathbf{X}_r = \left(\mathbf{x}_{r,1}, \ldots, \mathbf{x}_{r,j}, \ldots, \mathbf{x}_{r,M}\right)$. Здесь $j \le M$ – номер цикла наблюдения над r-й популяцией. И пусть $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\right)$ – аналогичная по структуре выборка объема M_x из речевого сигнала $\mathbf{x}$ с неизвестным распределением $\mathbf{P}_x \subset \{\mathbf{P}_r\}$ в пределах заданного множества альтернатив. Задача распознавания такого сигнала на множестве фонем $\{\mathbf{x}_r\}$ сводится к проверке на максимум функции правдоподобия объединенной выборки наблюдений $(\mathbf{X}, \mathbf{X}_\nu, \mathbf{X}_r)$ для всех $\nu, r = \overline{1, R}$. При общих условиях [17, с. 497] решение в пользу гипотезы $H_\nu, \nu \le R$, принимается по признаку выполнения R соотношений

$$\frac{\sup\limits_{\mathbf{K}_\nu}\left[p(\mathbf{X}/H_\nu)p(\mathbf{X}_\nu)\right]\sup\limits_{\mathbf{K}_r}\left[p(\mathbf{X}_r)\right]}{\sup\limits_{\mathbf{K}_r}\left[p(\mathbf{X}/H_r)p(\mathbf{X}_r)\right]\sup\limits_{\mathbf{K}_\nu}\left[p(\mathbf{X}_\nu)\right]} > 1, \quad \nu \ne r \le R. \quad (4)$$

Здесь $p(\mathbf{X}/H_r)$ – функция правдоподобия выборки $\mathbf{X}$ при справедливости гипотезы H_r; $p(\mathbf{X}_r)$ – функция правдоподобия r-й классифицированной выборки; символ sup обозначает верхнюю границу функции на множестве допустимых для каждой фонемы вариантов АКМ. При учете независимости наблюдений $\{\mathbf{x}_{r,j}\}_1^M$ в совокупности и в соответствии с известной методикой вычислений [12] запишем систему следующих равенств:

$$\ln p(\mathbf{X}/H_r) = \\ = -\frac{M_x}{2}\left[\ln|\mathbf{K}_r| + \text{tr}\left(\mathbf{S}_x\mathbf{K}_r^{-1}\right) + n\ln(2\pi)\right], \quad (5)$$

$$\ln p(\mathbf{X}_r) = \\ = -\frac{M}{2}\left[\ln|\mathbf{K}_r| + \text{tr}\left(\mathbf{S}_r\mathbf{K}_r^{-1}\right) + n\ln(2\pi)\right], \quad r = \overline{1, R}, \quad (6)$$

где $\mathbf{S}_r \triangleq M^{-1}\sum_{j=1}^{M}\mathbf{x}_{r,j}\mathbf{x}_{r,j}^{\mathrm{T}}$ – оценка максимального правдоподобия для АКМ $\mathbf{K}_r$ по выборке $\mathbf{X}_r$ объема M; $|\mathbf{K}_r|$ – определитель матрицы. Без нарушения общности формулировки задачи далее будем полагать [17], что объемы наблюдений в рабочем режиме и в режиме настройки (обучения) системы АОРР равны друг другу, т.е. $M_x = M$. Путем ряда вычислений имеем

$$\sup_{\mathbf{K}_r}\ln p(\mathbf{X}_r) = -\frac{M}{2}\left[\ln|\mathbf{S}_r| + nc\right],\quad r=\overline{1,R}, \qquad (7)$$

где $c = \ln(2\pi)+1 = \mathrm{const}$. Здесь учтено [12], что верхняя граница функций (5) и (6) достигается при равенстве АКМ $\mathbf{K}_r = \mathbf{S}_r$ – строго в соответствии с принципом максимума правдоподобия [17]. Аналогично, для всех других величин из выражения (4) получаем

$$\sup_{\mathbf{K}_r}\left[\ln p(\mathbf{X}|H_r)\,p(\mathbf{X}_r)\right] = -\frac{M}{2}\times$$
$$\times\left[2(\ln|\mathbf{S}_{xr}| + n\ln(2\pi)) + \mathrm{tr}(\mathbf{S}_x\mathbf{S}_{xr}^{-1}) + \mathrm{tr}(\mathbf{S}_r\mathbf{S}_{xr}^{-1})\right] =$$
$$= -M\left[\ln|\mathbf{S}_{xr}| + nc\right],\quad r=\overline{1,R}, \qquad (8)$$

где $\mathbf{S}_{xr} = 0.5(\mathbf{S}_x + \mathbf{S}_r)$ – оценка максимального правдоподобия для АКМ речевого сигнала по объединенной выборке наблюдений $\{\mathbf{X},\mathbf{X}_r\}$ суммарного объема $2M$. Подставляя выражения (7) и (8) в (4), после несложных преобразований получим искомый алгоритм распознавания сигнала $\mathbf{x}$ как ν-й фонемы при выполнении условия

$$M\left[\ln|\mathbf{S}_{x\nu}| - \ln|\mathbf{S}_{xr}| - 0.5\ln|\mathbf{S}_\nu| + 0.5\ln|\mathbf{S}_r|\right] < 0.$$

Или, в эквивалентном виде, можно записать

$$\rho_{x,x\nu} + \rho_{\nu,x\nu} < \rho_{x,xr} + \rho_{r,xr}. \qquad (9)$$

Здесь

$$\left.\begin{aligned}\rho_{x,xr} &\triangleq 0.5\left[\mathrm{tr}\left(\mathbf{S}_x\mathbf{S}_{xr}^{-1}\right) - \ln|\mathbf{S}_x| + \ln|\mathbf{S}_{xr}| - n\right]\\ \rho_{r,xr} &\triangleq 0.5\left[\mathrm{tr}\left(\mathbf{S}_r\mathbf{S}_{xr}^{-1}\right) - \ln|\mathbf{S}_r| + \ln|\mathbf{S}_{xr}| - n\right]\end{aligned}\right\}, \qquad (10)$$

– ВИР двух гауссовских распределений вероятностей с АКМ $\mathbf{S}_x$ и $\mathbf{S}_r$ соответственно относительно гипотетического гауссовского распределения с АКМ, равной $\mathbf{S}_{xr}$. Следуя критерию отношения правдоподобия, из (9) по индукции получим оптимальный алгоритм принятия решений

$$H_\nu : \rho_{x,xr} + \rho_{r,xr}\big|_{r=\nu} = \min \qquad (11)$$

в задаче АОРР (4). Набор оптимальных решающих статистик имеет в данном случае значительно более сложный вид по сравнению с его первоначальным вариантом (2), (3). Понятно, что это "плата" за априорную неопределенность в отношении статистических свойств фонетической базы данных диктора $\{\mathbf{x}_r\}$. В таком случае возникает закономерный вопрос о достигаемом выигрыше по эффективности оптимального алгоритма АОРР (11) по сравнению с алгоритмами на основе ДИС (3) адаптивного типа [16]. Ответом на него служат результаты дальнейшего исследования.

3. АСИМПТОТИЧЕСКИ ОПТИМАЛЬНЫЙ АЛГОРИТМ

При учете состоятельности оценок максимального правдоподобия [17] в асимптотике, когда объем выборки наблюдений $M_x \to \infty$, и в предположении о справедливости гипотезы H_ν можно записать $\mathbf{S}_\nu \to \mathbf{K}_\nu$, $\mathbf{S}_x \to \mathbf{K}_\nu$ и, следовательно, $\mathbf{S}_{x\nu} \to \mathbf{K}_\nu$, где символ $\to$ означает сходимость "*почти наверное*" (п.н.) или с вероятностью 1. На основании сказанного и равенств (10) приходим к двум очевидным импликациям для каждого слагаемого из левой части оптимального алгоритма (11):

$$\rho_{x,x\nu}\big|_{H_\nu} \xrightarrow{\text{п.н.}} 0 \Rightarrow \rho_{x,\nu}\big|_{H_\nu} \xrightarrow{\text{п.н.}} 0,$$
$$\rho_{\nu,x\nu}\big|_{H_\nu} \xrightarrow{\text{п.н.}} 0 \Rightarrow \rho_{\nu,x}\big|_{H_\nu} \xrightarrow{\text{п.н.}} 0.$$

Тем самым доказано следующее теоретическое положение.

Утверждение. Асимптотически оптимальное правило принятия решений в задаче АОРР общего вида (4) определяется выражением

$$H_\nu : \rho_{x,r} + \rho_{r,x}\big|_{r=\nu} = \min. \qquad (12)$$

Здесь сумма двух ВИР противоположной направленности определяет симметричную форму информационного рассогласования Кульбака–Лейблера [13, с. 16] между наблюдаемым сигналом $\mathbf{x}$ и его эталоном $\mathbf{x}_r$. Из алгоритма (12) согласно определению (3) получим выражение для асимптотически оптимальной решающей статистики в частотной области:

$$\rho_{x,r} = (2F)^{-1}\times$$
$$\times\int_{-F/2}^{F/2}\left(G_x(f)/G_r(f) + G_r(f)/G_x(f) - 2\right)\mathrm{d}f. \qquad (13)$$

Это известная [18] симметричная форма ДИС, или COSH-расстояние, для СПМ рассматриваемой пары сигналов. Ее практическая реализация в ИТВР хорошо изучена [2, 3]. Она основывается на методе обеляющего фильтра [10, 14] и на авторегрессионной модели речевого сигнала

$$G_x(f) = F^{-1}\sigma^2\left|1 + \sum_{m=1}^{p}a_x(m)\exp\left(-j\pi mf/F\right)\right|^2 \qquad (14)$$

конечного порядка p = 10...20, заданной своим вектором коэффициентов линейного предсказания $a_x(m)$, $m=\overline{1,p}$, и дисперсией порождающего

шума σ^2 [15]. Выражения (12)–(14) в совокупности и определяют искомый алгоритм АОРР в условиях априорной неопределенности в отношении спектрально-корреляционных свойств и характеристик речевого сигнала.

Как видим, синтезированный алгоритм в принципиальном отношении отличается от своего классического аналога (1)–(3). Решение в нем принимается с использованием не одной, а одновременно двух ДИС противоположной направленности для каждой альтернативы сигнала $\mathbf{x}_r$. Данный факт имеет решающее значение с точки зрения качества АОРР. Об этом свидетельствуют результаты проведенного автором статьи эксперимента.

4. ПРОГРАММА И РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

Качество АОРР на фонетическом уровне обработки речевого сигнала определяется в ИТВР [20] точностью и надежностью распознавания фонем в пределах речевого фрейма При этом точность алгоритма может быть охарактеризована [19] вероятностью $\alpha_{\nu\to r}$ ошибки первого рода или перепутывания ν-й фонемы с r-й, где $r \neq \nu \le R$. А его надежность – вероятностью β_ν ошибки второго рода или пропуска ν-й фонемы ввиду ее недостаточно четкого произнесения данным диктором. Вероятность $\alpha_{\nu\to r}$, в свою очередь, определяется [12] соответствующей ВИР $\rho_{\nu,r} = \rho(\mathbf{x}_\nu/\mathbf{x}_r)$ между фонетическими эталонами $\mathbf{x}_\nu$ и $\mathbf{x}_r$, в частности, по формуле ДИС (3), если обработка речевого сигнала $\mathbf{x}$ осуществляется в частотной области. Чем больше ВИР $\rho_{\nu,r}$ между фонемами в речи диктора, тем меньше вероятность их перепутывания. Аналогично, вероятность β_ν описывается известным выражением [19, 20] через ВИР аллофонов $\rho_{\nu j} = \rho(\mathbf{x}_{\nu j}/\mathbf{x}_\nu)$, $j = \overline{1, N}$, в пределах ν-го фонетического кластера $\{\mathbf{x}_{\nu,j}\}$ конечного объема N. Ее эталонный аллофон $\mathbf{x}_\nu$ в ИТВР определяют [11] как центр данного кластера в информационной метрике Кульбака–Лейблера (2). А среднее значение ВИР в пределах кластера $\rho_\nu \triangleq N^{-1}\sum_j \rho_{\nu j}$ служит рабочей характеристикой [11, 14] алгоритма АОРР в отношении ν-й фонемы. Величина ρ_ν зависит от диктора, а также от его функционального состояния в момент речеобразования [20]. Чем меньше эта величина, тем меньше вероятность ошибки второго рода β_ν. В таком случае отношение двух рассмотренных ВИР $\rho_{\nu,r}/\rho_\nu$ – это характеристика различимости двух соответствующих фонем при учете возможных ошибок их распознавания одновременно и первого, и второго рода. Как следствие, относительная величина

$$\mu_\nu = \frac{\min\limits_{r\le R}\rho_{\nu,r}}{\rho_\nu} \qquad (15)$$

может служить показателем эффективности распознавания ν-й фонемы в расчете на наихудший вариант фонемы $\mathbf{x}_r$ из множества ее альтернатив [19]. Чем больше величина μ_ν, тем выше эффективность алгоритма в расчете на конкретную фонему $\mathbf{x}_\nu$. Тогда на множестве из R разных фонем в речи диктора в качестве обобщенной характеристики алгоритма АОРР будем иметь векторный показатель эффективности $\{\mu_\nu\}_1^R$. Программа проведенного далее эксперимента предполагала сравнение по данному показателю двух конкурирующих алгоритмов обработки речевого сигнала: на основе классической формы ДИС (3) и на основе COSH-расстояния (13).

Объектом экспериментального исследования служили сигналы русских гласных фонем (случай $R = 6$) в произнесении контрольного диктора (в данном случае – автора статьи) как наиболее информативные в коммуникативном смысле [7, 21]. Методика исследования предполагала раздельную статистическую оценку числителя и знаменателя из выражения (15) по формуле средней выборочной величины. Для этого использовали специальную компьютерную программу “Phoneme Training”, которая находится в открытом доступе[2]. Скриншот ее главного окна показан на рис. 1.

В правой части экрана (см. рис. 1) отображается график СПМ сигнала произнесенной гласной фонемы. Его длительность $T_x = N\tau$ составляла в эксперименте не менее 2...4 с. При этом объем контрольной выборки $N = 200...400$ варьировался в широких пределах – в расчете на статистическое усреднение результатов АОРР на множестве речевых фреймов. Их длительность $\tau = 10$ мс была ограничена по времени условиями обеспечения стационарности речевого сигнала на интервале его автоматической обработки. В программе этот сигнал сначала автоматически редактировался – для отсечения переходных процессов в его начале и конце – и только после этого членился на множество (N-последовательность) фреймов данных $\mathbf{x}_j$.

Все основные параметры программы были установлены равными их апробированным ранее значениям [2, 3]: $F = 8$ кГц, $n = 80$, $p = 20$ и $\sigma^2 = 1$. А для определения вектора коэффициентов авторегрессии $\{a_x(m)\}_1^{20}$ из выражения (14) по выборке $\mathbf{x}_j$ была применена высокоскоростная рекур-

[2] См. ссылку в Интернете https://sites.google.com/site/frompldcreators/.

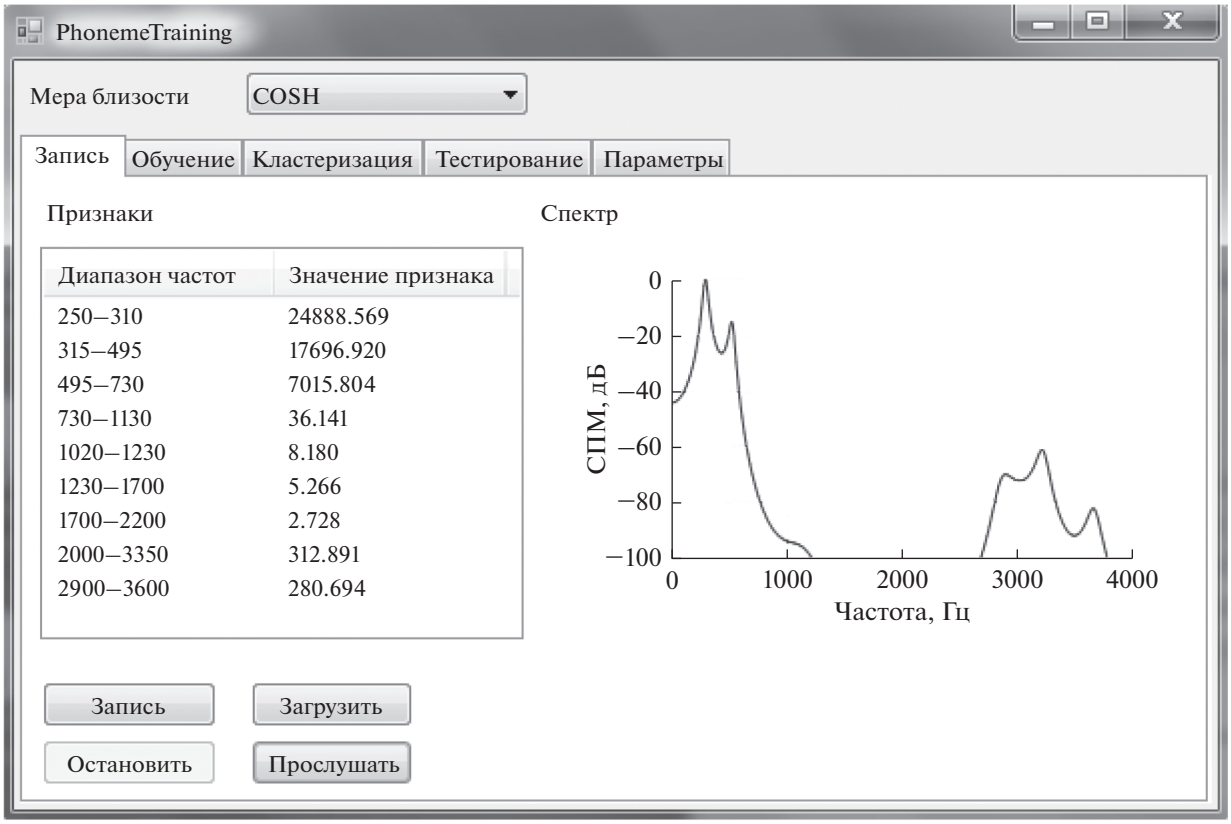

Рис. 1. Скриншот главного окна компьютерной программы в процессе обработки сигнала фонемы “у”.

рентная процедура Берга–Левинсона [15]. Все вычисления при этом производились на современном персональном компьютере с использованием стандартного программного обеспечения.

Формирование фонетической базы данных контрольного диктора в программе осуществлялось по известной методике [3]: в режиме “Кластеризация” на множестве отрезков сигнала $\{\mathbf{x}_{r,j}\}$ каждой r-й фонемы. Одновременно была рассчитана и средняя ВИР ρ_r в пределах каждого фонетического кластера. Точность полученных оценок на уровне значимости γ может быть охарактеризована величиной их относительной погрешности [22] $\delta = z_{1-\gamma/2}/\sqrt{N}$, где $z_{1-\gamma/2}$ – квантиль порядка $q = 1-\gamma/2$. Например, при $\gamma = 0.05$ и $N = 400$ будем иметь [17] $z_{1-\gamma/2} = z_{0.975} \approx 1.96$ и $\delta \le 10\%$, что представляется [20, 21] вполне приемлемым результатом в условиях малых выборок наблюдений.

В продолжение эксперимента была поэлементно построена матрица ВИР $\|\rho_{v,r}\|$ размером 6×6 на множестве эталонов $\{\mathbf{x}_r\}_1^6$ гласных фонем от контрольного диктора. Для этого программа была переведена в режим “Обучение”. В качестве примера на рис. 2 показан скриншот окна программы при работе в этом режиме с эталоном фонемы “ы”. Здесь в колонке “Расстояние” перечислены все элементы из соответствующей строки матрицы ВИР. По матрице ВИР и вектору средних значений $\{\rho_v\}$ были рассчитаны согласно выражению (15) векторные показатели эффективности алгоритмов АОРР на основе ДИС и COSH-расстояния: $\{\mu_v\}$ и $\{\tilde{\mu}_v\}$ соответственно. Полученные результаты представлены в табл. 1.

Из сравнения данных, полученных на основе двух методов, видно, что использование COSH-расстояния (13) в роли решающей статистики асимптотически оптимального алгоритма (11) позволяет существенно повысить эффективность системы АОРР в отношении большинства гласных

Таблица 1. Показатели эффективности $\{\mu_v\}$ двух алгоритмов АОРР

Алгоритм	Фонемы					
	“а”	“и”	“о”	“у”	“ы”	“э”
	1	2	3	4	5	6
ДИС	15.95	24.05	25.41	31.61	25.48	47.78
COSH	21.34	45.43	28.28	41.14	44.44	51.52

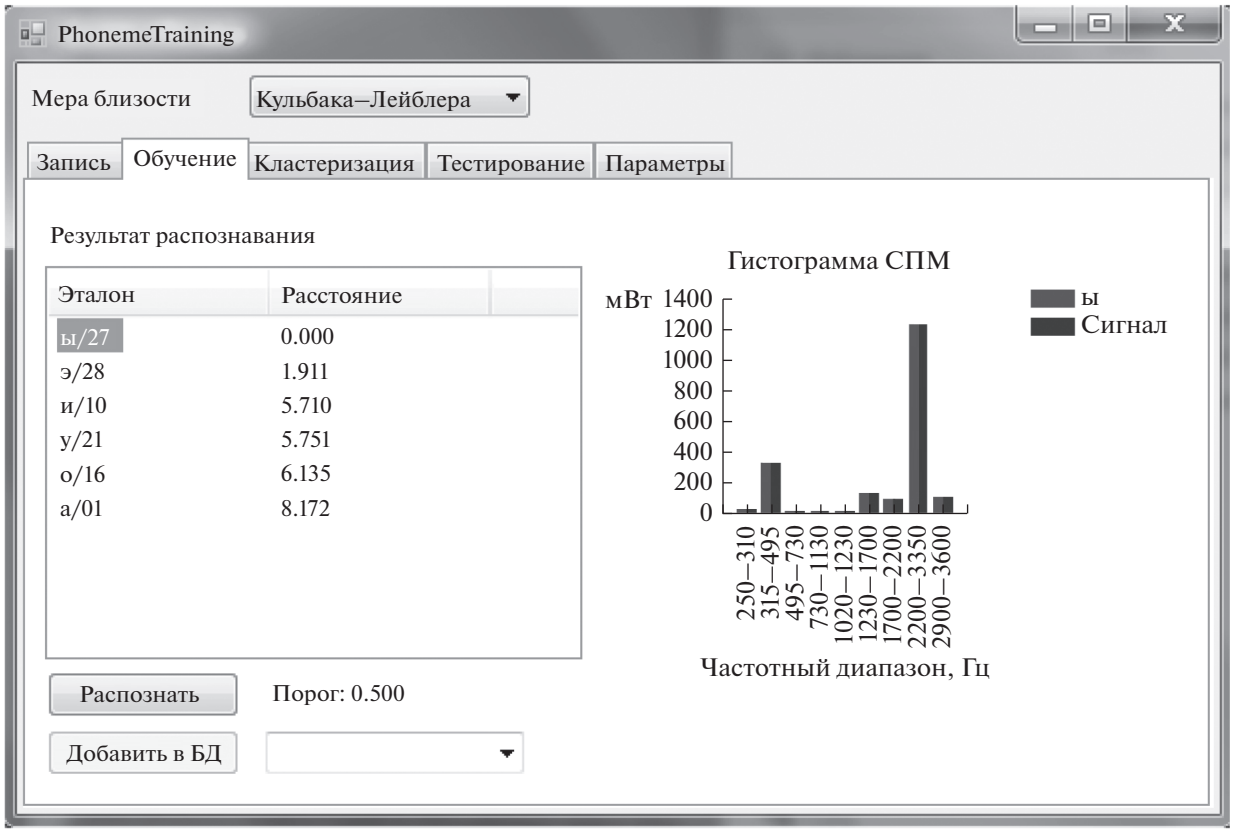

Рис. 2. Скриншот рабочего окна программы в режиме “Обучение” при работе с эталоном фонемы “ы”.

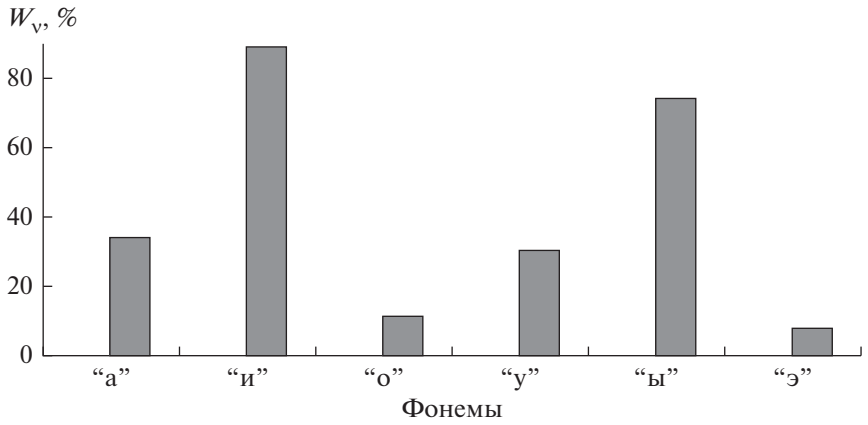

Рис. 3. Гистограмма выигрыша по эффективности W_v, %.

фонем. Величина достигаемого в данном случае выигрыша (рис. 3) может быть рассчитана по интуитивно понятной формуле

$$W_v = \frac{\tilde{\mu}_v - \mu_v}{\mu_v} \times 100, \%. \quad (16)$$

Отметим, что максимум выигрыша (16) обеспечивается при обработке наиболее сложно распознаваемых на практике [21] гласных фонем, таких как “и” и “ы”. Напротив, выигрыш минимален для фонемы “э”, которая характеризуется наиболее высоким показателем эффективности АОРР

(см. табл. 1) согласно его определению (15). Подробнее данный эффект рассмотрен далее.

5. ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Вид матрицы $\|\rho_{\nu,r}\|$ зависит от акустических свойств речевого сигнала конкретного диктора. Для иллюстрации представлена матрица ВИР шести гласных фонем от контрольного диктора (табл. 2). В ней каждая фонема обозначена соответствующим номером из табл. 1, а элементом $\rho_{\nu,r}$ на пересечении ее ν-й строки и r-го столбца служит ВИР (3) для пары фонетических эталонов $(\mathbf{x}_\nu, \mathbf{x}_r)$. Для сравнения на главной диагонали матрицы приведены средние значения ВИР $\{\rho_\nu\}$ (выделены серым фоном) в пределах соответствующего фонетического кластера $\{\mathbf{x}_{\nu,j}\}$. Отметим принципиально асимметричный вид данной матрицы. Это следствие известного свойства асимметрии ДИС [9] согласно ее определению (3). Причем различия в элементах $\rho_{\nu,r}$ и $\rho_{r,\nu}$ здесь достигают весьма значительной величины: десятки раз и более. В нашем примере максимумом подобных различий характеризуются такие фонетические пары, как “а”–“о”, “а”–“у”, “о”–“и” и др. Наибольшего внимания заслуживает пара “ы”–“э”, для которой одна из ВИР противоположной направленности, а именно: $\rho_{5,6}$ не превышает пороговой величины [19] $\rho_0 = 1.5...2$. В результате для этой пары резко возрастает вероятность перепутывания фонем между собой. Однако проблема носит принципиально односторонний характер. В частности, если вероятность перепутывания “ы” с “э” для данного диктора весьма ощутима [21], то вероятностью противоположного события можно пренебречь. Но справедливо это только для алгоритма АОРР на основе ДИС (3). В асимптотически оптимальном алгоритме (11) на основе COSH-расстояния (13) роль рабочей характеристики выполняет полусумма ВИР противоположной направленности $0.5(\rho_{\nu,r} + \rho_{r,\nu})$, что практически исключает [18] ошибки, подобные рассмотренной выше. Действительно, в нашем примере с парой фонем “ы”–“э” имеем $0.5(\rho_{5,6} + \rho_{6,5}) \approx 3.15 > \rho_0$, а это гарантия [12] достаточно высокой точности и надежности их автоматического распознавания. Таким образом, именно асимметрия ДИС служит обоснованием достигаемого согласно выражению (16) выигрыша асимптотически оптимального алгоритма АОРР (12)–(14) по эффективности. Аналогичный эффект был исследован, правда с иных позиций, в работе [20].

Таблица 2. Матрица ВИР $\|\rho_{\nu,r}\|$

ν	r					
	1	2	3	4	5	6
1	0.15	30.70	3.18	3.09	5.04	2.44
2	7.84	0.10	21.34	27.34	2.29	4.52
3	75.94	546.90	0.09	2.31	92.70	25.45
4	133.85	199.88	2.21	0.07	33.88	24.39
5	8.17	5.71	6.14	5.75	0.08	1.91
6	3.54	42.83	14.82	13.32	4.40	0.07

Наглядной иллюстрацией к сказанному служат графики СПМ проблемной пары фонем на рис. 4. Здесь хорошо видна их особенность: две эти СПМ похожи между собой, но для фонемы “э” спектр мощности “богаче” на одну моду в области средних частот. Как результат, при обработке речевого сигнала методом обеляющего фильтра [10] практически исключается возможность перепутывания фонемы “э” с фонемой “ы”. Однако обратное перепутывание в процессе АОРР на основе односторонней ДИС (3) характеризуется весьма существенной вероятностью $\alpha_{r\to\nu}$.

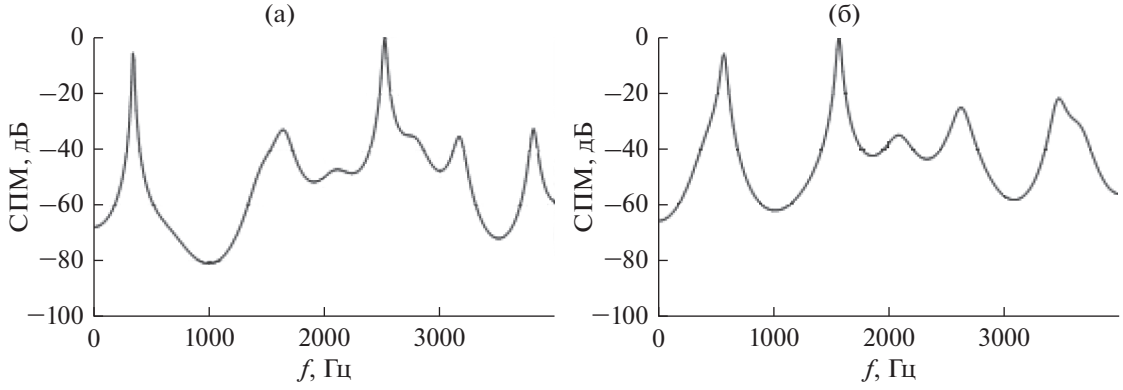

Рис. 4. Скриншоты фрагментов главного окна программы с графиками СПМ фонемы “ы” (а) и фонемы “э” (б).

Ситуация исправляется в принципиальном отношении при применении COSH-расстояния (13).

ЗАКЛЮЧЕНИЕ

Таким образом, проведенное исследование дало строгое теоретико-информационное обоснование симметричной формы ДИС в роли асимптотически оптимальной решающей статистики в задачах автоматической обработки и распознавания речевых сигналов. Полученные результаты позволят исследователям и разработчикам современных речевых систем и технологий находить наиболее эффективные технические решения острейшей в данной области науки и техники [1–3] проблемы: защиты от интенсивных акустических помех [4–6], когда известные методы обработки сигналов зачастую не обеспечивают требуемой точности и надежности распознавания отдельных фонетических единиц.

СПИСОК ЛИТЕРАТУРЫ

1. *Zhow J., Zheng W., Wang Q., Zhao L.* // Chinese J. Acoust. 2014. T. 33. № 3. C. 312.
2. *Савченко В.В.* // РЭ. 2017. Т. 62. № 7. С. 681.
3. *Savchenko V.V.* // Radiophysics and Quantum Electron. 2015. V. 58. № 5. P. 373.
4. *Schuster M.* // Lecture Notes in Computer Sci. 2010. V. 6230. P. 8.
5. *Савченко В.В.* // РЭ. 2016. Т. 61. № 12. С. 1196.
6. *Benesty J., Sondhi M.M., Huang Y.* // Springer handbook of speech processing. N.Y.: Springer, 2008. Pt B.
7. *Савченко В.В., Савченко А.В.* // РЭ. 2016. Т. 61. № 4. С. 373.
8. *Chen G., Koh S.N., Soon I.Y.* // Signal Processing. 2003. V. 83. P. 1445.
9. *Gray R.M., Buzo A., Gray A.H., Matsuyama Y.* // IEEE Trans. 1980. V. ASSP-28. № 4. P. 367.
10. *Савченко В.В.* // РЭ. 2005. Т. 50. № 3. С. 309.
11. *Савченко В.В.* // Изв. вузов. Радиоэлектроника. 2007. № 6. С. 3.
12. *Савченко В.В.* // Изв. вузов. Радиоэлектроника. 2012. № 2. С. 47.
13. *Kullback S.* Information Theory and Statistics. N.Y.: Dover Publ., 1997.
14. *Савченко В.В.* // РЭ. 1997. Т. 42. № 4. С. 426.
15. *Marple S.L.* Digital Spectral Analysis. Englewood Cliffs, NJ: Prentice Hall, 1987.
16. *Savchenko A.V.* // Lecture Notes in Computer Sci. 2014. V. 8509. P. 638.
17. *Боровков А.А.* Математическая статистика. СПб.: Лань, 2010.
18. *Wei B., Gibson J.* // Proc. IEEE Digital Signal Processing Workshop. Hunt, Texas, 2000. P. 3.
19. *Савченко В.В.* // Электросвязь. 2017. № 12. С. 22.
20. *Савченко В.В.* // РЭ. 2018. Т. 63. № 1. С. 60.
21. *Конев А.А., Мещеряков Р.В., Ходашинский И.А.* // VI Междисциплинарный семинар “Анализ разговорной русской речи” (АРЗ-2012). СПб.: Изд-во СПГУ, 2012. С. 35.
22. *Савченко В.В.* // Научные ведомости Белгород. гос. ун-та. Сер. Экономика. Информатика. 2015. Т. 33/1. № 1. С. 74.