РАДИОТЕХНИКА И ЭЛЕКТРОНИКА, 2020, том 65, № 11, с. 1101–1108

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ СИГНАЛОВ

УДК 621.391:004.934

КРИТЕРИЙ ГАРАНТИРОВАННОГО УРОВНЯ ЗНАЧИМОСТИ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА

© 2020 г. В. В. Савченко[a], *, А. В. Савченко[b], **

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

[b] *Национальный исследовательский университет "Высшая школа экономики", Лаборатория алгоритмов и технологий анализа сетевых структур, ул. Б. Печерская, 25, Нижний Новгород, 603155 Российская Федерация*

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

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

Поступила в редакцию 14.02.2019 г.
После доработки 07.02.2020 г.
Принята к публикации 20.04.2020 г.

Рассмотрена задача автоматического выделения из речевого сигнала его фонетических единиц в условиях априорной неопределенности относительно их спектрального состава и корреляционных свойств. На основе теоретико-информационного подхода разработан критерий гарантированного уровня значимости. Рассмотрен пример его практического применения, поставлен и проведен натурный эксперимент. Показано, что благодаря предложенному критерию гарантируется стабильный уровень значимости при обработке речевых фреймов малой длительности.

DOI: 10.31857/S0033849420110157

ВВЕДЕНИЕ

Под сегментацией сигнала в задачах автоматического распознавания речи (АРР) традиционно понимают [1, 2] ее фонемную или, иными словами [3], фонологическую разновидность, целью которой является on-line-членение речевого потока на последовательность минимальных (не делимых далее) речевых единиц типа фонем и их аллофонов. Это важнейшая составная часть обработки речевого сигнала в системах самого разного назначения [4–6]: от голосового управления и идентификации дикторов до речевой аналитики, и биометрии, которая между тем зачастую недооценивается специалистами. Причина сказанного кроется в самом понятии фонологической сегментации, предшествующей этапу распознавания (парадигматической идентификации [3]) вычлененных сегментов сигнала в рамках "отложенной" [7] сегментации речи. Так, например, в работах [8, 9] применен простейший способ фонологической сегментации: членение речевого сигнала на речевые фреймы (отрезки сигнала) предельно малой длительности $\tau = 10...20$ мс, которая согласована с периодом основного тона устной речи типичного диктора [9]. Однако в этом случае возникает [10] острая проблема малых выборок наблюдений и вслед за ней обостряется проблема множественных сравнений [11]. Как следствие, приходится констатировать [7, 12], что применительно к русской слитной речи с большим словарем указанная задача до настоящего времени не решена совсем или решена недостаточно эффективно. Между тем, как это показано в работах [13, 14] на ряде примеров из практики, при применении сегментации речевого сигнала с объединением однородных фреймов в однофонемные сегменты речи удается в значительной степени преодолеть проблему малых выборок, а вслед за ней – и множественных сравнений в задачах АРР. Поэтому можно утверждать [15–17], что полноценная фонологическая сегментация является в настоящее время наиболее перспективным способом повышения эффективности АРР на стадии первичной обработки речевого сигнала [7]. Первостепенное значение при этом имеет вопрос о выборе критерия сегментации [3]. Поэтому актуальность темы проведенного далее исследования представляется очевидной.

В основу предложенного в статье критерия положен принцип его гарантированного уровня значимости в задаче обнаружения "разладки" случайного сигнала [18–21] на интервале длительностью в один речевой фрейм. В отличие от известных критериев [13–17] он напрямую не связан с понятием случайной погрешности статистических оценок параметров распределений и

нацелен на применение в условиях априорной неопределенности в отношении тонкой структуры речевого сигнала [22].

1. ПОСТАНОВКА ЗАДАЧИ

Следуя статистической теории "разладки" [19], воспользуемся универсальной [9–11] гауссовой аппроксимацией $P(X_k) = \mathrm{Norm}_n(\mathbf{R}_k)$ многомерного (n-мерного) закона распределения наблюдаемого сигнала $x(t)$ в пределах одного (текущего) речевого фрейма X_k фиксированной длительности $\tau = \mathrm{const}$, где $k = 1, 2, \ldots$. Здесь $\mathbf{R}_k \triangleq \mathbf{E}\left(\mathbf{x}_k\mathbf{x}_k^T\right)$ – автокорреляционная $(n\times n)$-матрица (АКМ) речевого сигнала, который предполагается предварительно центрированным; $\mathbf{x}_k$ – n-вектор (столбец) его последовательных отсчетов (символами $\mathbf{E}(\cdot)$, $\triangleq$ и $(\cdot)^T$ обозначены соответственно математическое ожидание, равенство по определению и операция транспонирования векторов). Задача формулируется в терминах проверки статистических гипотез

$$\left.\begin{array}{l} H_0 : \mathbf{R}_k = \mathbf{R}_{k-1} \triangleq \mathbf{R} \\ H_1 : \mathbf{R}_k \neq \mathbf{R}_{k-1} \end{array}\right\}, \quad k = 1, 2, \ldots, \qquad (1)$$

о равенстве друг другу АКМ речевого сигнала в двух соседних фреймах X_{k-1} и X_k. Она решается пошагово. Здесь k – номер шага с инициализацией в виде равенства $k = 1$. По результатам решения задачи (1) на каждом очередном шаге k текущий речевой фрейм X_k либо объединяется с предыдущим фреймом X_{k-1} в один однородный сегмент речевого сигнала, либо, напротив, обособляется в качестве первого фрейма очередного сегмента в речи диктора. Во втором случае номер k текущего речевого фрейма вновь устанавливается равным единице.

Задача состоит, таким образом, в последовательной – от фрейма к фрейму – проверке статистических гипотез (1) в пределах интервала наблюдения над речевым сигналом $x(t)$. При этом инициализацией вычислительной процедуры (1) может служить равенство $\mathbf{R}_0 = \mathrm{diag}_n\left(\sigma_0^2\right)$, где символом $\mathrm{diag}_n(\cdot)$ обозначена диагональная $(n\times n)$-матрица с дисперсией σ_0^2 фонового (из речевых пауз) шума на главной диагонали.

В условиях априорной неопределенности, когда матрицы $\mathbf{R}_k$ и $\mathbf{R}_{k-1}$ заранее неизвестны, воспользуемся их оценками максимального правдоподобия по формуле корреляционного выборочного момента [22–24]

$$\mathbf{S}_j = M^{-1}\sum_{i=1}^{M}\mathbf{x}_{j,i}\mathbf{x}_{j,i}^T, \quad j = k-1,\ k, \qquad (2)$$

где $\mathbf{x}_{k,i}$ – i-й (парциальный) n-вектор последовательных отсчетов речевого сигнала; $M \triangleq [N/n]$ (целая часть числа) – количество непересекающихся парциальных векторов в пределах одного (наблюдаемого) фрейма; $N = F\tau$ – суммарный объем выборки из речевого сигнала на интервале в один фрейм; F – частота его дискретизации. При этом размерность векторов $\mathbf{x}_{k,i}$, $i \le M$, определяется наблюдателем в зависимости от полосы частот $[F_{\min};\ F_{\max}]$ в спектре речевого сигнала [4]: $n = 0.5F_{\max}/F_{\min} = 0.25F/F_{\min}$. Так, при частоте дискретизации $F = 8$ кГц (согласована с полосой частот стандартного телефонного канала связи [11]), $F_{\min.} = (100...200)$ Гц и длительности фрейма $\tau = 10$ мс будем иметь $n = 10...20$, $N = 80$ и, следовательно, получаем $M = (4...8)$ парциальных выборок для вычислений матрицы $\mathbf{S}_j$. А это явный признак остроты проблемы малых выборок наблюдений [10]. Поэтому воспользуемся для решения задачи (1) асимптотически минимаксным критерием отношения правдоподобия с гарантированным уровнем значимости [22].

2. СИНТЕЗ АЛГОРИТМА

Определим в рамках указанного критерия критическую область n-мерного выборочного пространства согласно решающему правилу общего вида [19]:

$$W: \ \lambda(X_0) \triangleq \frac{\sup\limits_{\mathbf{R}_{k-1},\,\mathbf{R}_k} p(X_0|H_1)}{\sup\limits_{\mathbf{R}_0} p(X_0|H_0)} > \lambda_0, \qquad (3)$$

где $p(X_0|H_0) = p(X_{k-1}|H_0)p(X_k|H_0)$, $p(X_0|H_1) = p(X_{k-1}|H_1)p(X_k|H_1)$ – функции правдоподобия соответственно гипотез H_0 и H_1 для объединенной выборки наблюдений ($P(\cdot\,|\,\cdot)$ – условная вероятность случайного события, символом sup обозначена верхняя граница функции на множестве допустимых АКМ $\mathbf{R}_j$, $j = k-1,\ k$). Уровень значимости данного критерия $\alpha \triangleq P(W|H_0)$ регулируется выбором порогового уровня λ_0 в правой части выражения (3). Применительно к рассматриваемой задаче автоматической сегментации речи такая регулировка позволяет менять в широких пределах и при этом гарантировать выполнение требований наблюдателя [6] к степени однородности речевого сигнала в пределах каждого отдельного фонетического сегмента данных.

В условиях априорной неопределенности наблюдателю неизвестны распределения $p(X_j|H_1)$ и $p(X_j|H_0)$. Поэтому, следуя общесистемному принципу максимума энтропии [23–26], раскроем правило принятия решений (3) в расчете на

максимально неопределенный случай: статистической независимости выборок $\mathbf{x}_{j,\,i}$, $i \le M$, в совокупности. Для этого случая запишем систему равенств [5]

$$p(X_j|H_1) = \prod_{i=1}^{M} p(\mathbf{x}_{j,i}) = (2\pi)^{-nM/2} \left|\mathbf{R}_j\right|^{-M/2} \times$$
$$\times \exp\left(-0.5\sum_{i=1}^{M} \mathbf{x}_{j,i}^T \mathbf{R}_j^{-1} \mathbf{x}_{j,i}\right),$$
$$p(X_j|H_0) = (2\pi)^{-nM/2} \left|\mathbf{R}_j\right|^{-M/2} \times$$
$$\times \exp\left[-0.5\sum_{i=1}^{M} \mathbf{x}_{j,i}^T \mathbf{R}_j^{-1} \mathbf{x}_{j,i}\right], \quad j = k-1,\ k.$$

Или, после логарифмирования, будем иметь

$$\ln p\left(X_j|H_1\right) = -0.5M\left[\ln\left|\mathbf{R}_j\right| + \mathrm{tr}\left(\mathbf{S}_j\mathbf{R}_j^{-1}\right) + nc\right],$$
$$\ln p\left(X_j|H_0\right) = -0.5M\left[\ln\left|\mathbf{R}\right| + \mathrm{tr}\left(\mathbf{S}_0\mathbf{R}^{-1}\right) + nc\right],$$
$$j = k-1,\ k.$$

Здесь $\mathbf{S}_0 = 0.5\left(\mathbf{S}_{k-1} + \mathbf{S}_k\right)$ – оценка максимума правдоподобия для АКМ речевого сигнала по объединенной выборке наблюдений X_0, где $c = \ln(2\pi) = \mathrm{const}$ (символами $|\cdot|$ и $\mathrm{tr}(\cdot)$ обозначены соответственно определитель и след квадратной $(n \times n)$-матрицы). Путем несложных вычислений [19] отсюда получаем

$$\left.\begin{aligned} \ln\sup p\left(X_j|H_1\right) &= -0.5M\left[\ln\left|\mathbf{S}_j\right| + n(c+1)\right], \quad \forall j = k-1,\quad k, \\ \ln\sup p\left(X_0|W_0\right) &= -M\left[\ln\left|\mathbf{S}_0\right| + n(c+1)\right], \end{aligned}\right\}. \tag{4}$$

При этом было учтено [22], что на множестве допустимых ковариаций выборочных данных X_k и X_{k-1} верхняя граница функций правдоподобия достигается при выборе АКМ $\mathbf{R}_k$ и $\mathbf{R}_{k-1}$ равными их оценкам максимума правдоподобия $\mathbf{S}_k$ и $\mathbf{S}_{k-1}$ соответственно, если справедлива гипотеза H_1, и $\mathbf{R} = \mathbf{S}_0$ – в противном случае. Здесь все АКМ, как и их выборочные оценки, предполагаются неособенными. Проблема их обусловленности на практике преодолевается [4] путем оптимизации параметров n и M и применением современных вычислительных процедур корреляционно-спектрального анализа [23].

Полученные выражения (4) определяют общую формулировку для оптимальной решающей статистики вида

$$\tilde{\lambda}(X_0) \triangleq 0.5\left(2\ln\left|\mathbf{S}_0\right| - \ln\left|\mathbf{S}_k\right| - \ln\left|\mathbf{S}_{k-1}\right|\right) = \\ = 2H_n(\mathbf{S}_0) - H_n(\mathbf{S}_k) - H_n(\mathbf{S}_{k-1}), \tag{5}$$

где $H_n(\mathbf{S}_j) = 0.5\left(\ln\left|\mathbf{S}_j\right| + nc\right)$ – дифференциальная (по Шеннону) энтропия [24] n-мерного гауссова распределения вероятностей с АКМ, равной $\mathbf{S}_j$, $j = 0,\ k-1,\ k$. Решение здесь принимается по принципу допустимых различий между двумя эмпирическими распределениями, $\mathrm{Norm}_n(\mathbf{S}_{k-1})$ и $\mathrm{Norm}_n(\mathbf{S}_k)$, в теоретико-информационном смысле. В идеальном случае, когда выполняется система равенств $\mathbf{S}_0 = \mathbf{S}_{k-1} = \mathbf{S}_k$, имеем равенство $\tilde{\lambda}(X_0) = 0$. Но это, повторяем, только в идеальном случае. В реальности будем иметь $\mathbf{S}_{k-1} \ne \mathbf{S}_k$ и, следовательно, выполняется неравенство $\tilde{\lambda}(X_0) \ne 0$. Его характер уточняется в следующем теоретическом положении.

Утверждение. В условиях вывода равенства (5) выполняется соотношение $\tilde{\lambda}(X_0) \ge 0$.

Доказательство. Отталкиваясь от выражения (5), запишем

$$\tilde{\lambda}(X_0) = 0.5\left[\mathrm{tr}\left(\mathbf{S}_{k-1}\mathbf{S}_0^{-1}\right) - \ln\frac{\left|\mathbf{S}_{k-1}\right|}{\left|\mathbf{S}_0\right|} + \right.$$
$$\left. + \mathrm{tr}\left(\mathbf{S}_k\mathbf{S}_0^{-1}\right) - \ln\frac{\left|\mathbf{S}_k\right|}{\left|\mathbf{S}_0\right|} - 2n\right] = \Theta_{0/k-1} + \Theta_{0/k},$$

где

$$\Theta_{0/j} = 0.5\left[\mathrm{tr}\left(\mathbf{S}_j\mathbf{S}_0^{-1}\right) - \ln\frac{\left|\mathbf{S}_j\right|}{\left|\mathbf{S}_0\right|} - n\right]$$

– величина информационного рассогласования по Кульбаку–Лейблеру [24] двух гауссовых распределений с АКМ $\mathbf{S}_0$ и $\mathbf{S}_j$, $j = k-1,\ k$, обладающая свойством $\Theta_{0/j} \ge 0$ и с равенством нулю в случае равенства двух рассматриваемых АКМ друг другу. Отсюда вытекает справедливость сформулированного выше утверждения.

Следствие. Доказанное утверждение приводит к следующей импликации: $\tilde{\lambda}(X_0) \ge 0 \Rightarrow$ $\Rightarrow \{\ln|\mathbf{S}_0| \ge \ln|\mathbf{S}_{k-1}|\} \vee \{\ln|\mathbf{S}_0| \ge \ln|\mathbf{S}_k|\}$ – достоверное событие, где символом $\vee$ обозначен квантор объединения двух случайных событий, или их дизъюнкция.

На основании последнего утверждения и выражения (5) подоптимальное правило принятия решений в задаче проверки статистических гипотез (1) может быть представлено в виде

$$W: \ln\frac{\left|\mathbf{S}_0\right|}{\left|\mathbf{S}_{k-1}\right|} > \tilde{\lambda}_0 \vee \ln\frac{\left|\mathbf{S}_0\right|}{\left|\mathbf{S}_k\right|} > \tilde{\lambda}_0. \tag{6}$$

При этом учитываются оба возможных варианта проявления различий в эмпирических распределениях $\mathrm{Norm}_n(\mathbf{S}_{k-1})$ и $\mathrm{Norm}_n(\mathbf{S}_k)$: в сторону как

увеличения, так и уменьшения их энтропий $H_n(\mathbf{S}_k)$ и $H_n(\mathbf{S}_{k-1})$ по отношению к энтропии $H_n(\mathbf{S}_0)$ распределения объединенной выборки X_0. Пороговый уровень $\tilde{\lambda}_0 > 0$ служит здесь регулятором уровня значимости $\alpha(\delta_0)$. Указанная регулировка является существенным преимуществом решающего правила (6) по сравнению с его известными аналогами. Проиллюстрируем данное преимущество на конкретном примере практической реализации алгоритма (6) с использованием авторегрессионной модели речевого сигнала и математического аппарата авторегрессионного анализа [23, 25].

3. ПРИМЕР ПРАКТИЧЕСКОЙ РЕАЛИЗАЦИИ

Авторегрессионная модель (АР-модель) речевого сигнала [23]

$$x_j(t) = \sum_{i=1}^{q} a_q(i)x_j(t-i) + \eta_j(t), \quad j = k-1, \quad k, \quad (7)$$

в пределах j-го речевого фрейма X_j однозначно определяется своим вектором АР-коэффициентов $\{a_q(i),\ i = \overline{1,q}\}$ заданного порядка $q \le n$ и дисперсией $\sigma_j^2 = \text{const}$ порождающего процесса $\{\eta_j(t)\}$ типа белого гауссова шума в дискретном времени $t = 1,2,\ldots$. С одной стороны, модель (7) органично сочетается с голосовым механизмом человека (имеется в виду известная [4, 6] модель “акустической трубы”), с другой – существенно расширяет возможности программно-аппаратной реализации критерия (6). С указанной точки зрения представляет интерес асимптотическое равенство [25, 26]: $n^{-1}\ln\left|\mathbf{S}_j\right|_{n\to\infty} = \ln\sigma_j^2$. Величина σ_j^2 здесь характеризует [23] минимально достижимую дисперсию погрешности линейного предсказания случайного временного ряда (7) на один шаг в будущее. Теоретически данное равенство с достаточной степенью точности обусловлено соотношением $N \gg q$ [25].

В самом деле, порядок АР-модели на практике [10, 11] не превышает $q = 10\ldots20$, притом что размер АКМ речевого сигнала в задачах АРР ограничен только объемом $N = 80\ldots120$ речевого фрейма. Критерий (6) в таком случае может быть переписан следующим образом:

$$W:\ \ A_{k-1} : \ln\frac{\sigma_0^2}{\sigma_{k-1}^2} > \delta_0 \vee A_k : \ln\frac{\sigma_0^2}{\sigma_k^2} > \delta_0$$

– через объединение двух случайных событий A_{k-1} и A_k для принятия гипотезы H_1. Или, что эквивалентно, – в виде условия к максимуму отношения двух дисперсий

$$W:\ \max_{j=k-1;k}\left(\frac{\sigma_0^2}{\sigma_j^2}\right) > \tilde{\delta}_0 \quad (8)$$

погрешности линейного предсказания речевого сигнала в текущем времени $k = 1, 2, \ldots$ Здесь $\tilde{\delta}_0 > 1$ – пороговый уровень, а символом $(\cdot)$ обозначена выборочная оценка дисперсии погрешности

$$\varepsilon_j(t) = x(t) - \sum_{i=1}^{q} a_q(i)x(t-i), \quad t = 1,2,\ldots, \quad (9)$$

линейного предсказания q-го порядка в пределах j-го речевого фрейма.

Полученный результат (8), (9) – это известная [26] формулировка критерия проверки статистических гипотез о равенстве дисперсий откликов двух обеляющих фильтров на речевой сигнал. Их назначение – декорреляция сигнала $\varepsilon_j(t)$ на выходе. Указанная декорреляция достигается, если вектор АР-коэффициентов $\{a_q(i)\}$ фильтра (9) предварительно адаптирован под анализируемый сигнал $\{x(t)\}$ на интервале его наблюдения в один или два фрейма подряд (при равенстве $j = 0$). Для этого в теории авторегрессионного анализа разработан эффективный математический аппарат. В качестве примера можно привести высокоскоростную вычислительную процедуру Берга–Левинсона вида [23]

$$\left.\begin{aligned} a_m(i) &= a_{m-1}(i) + c_m a_{m-1}(m-i), \quad i = \overline{1,m};\\ c_m &= (N-m)^{-1}S_{m-1}^{-2}\sum_{t=m+1}^{N}\eta_{m-1}(t)\nu_{m-1}(t-1),\\ S_{m-1}^2 &= 0.5(N-m)^{-1}\sum_{t=m+1}^{N}\left[\eta_{m-1}^2(t) + \nu_{m-1}^2(t-1)\right];\\ \eta_m(t) &= \eta_{m-1}(t) - c_m\nu_{m-1}(t-1),\\ \nu_m(t) &= \nu_{m-1}(t-1) - c_m\eta_{m-1}(t), \quad t = 1,2,\ldots,N,\\ m &= \overline{1,q}, \end{aligned}\right\} \quad (10)$$

при ее инициализации системой равенств $\nu_0(t) = \eta_0(t-1) = x(t)$, $t = 1,2,\ldots$ Финальные значения данной рекурсии (10) при $m = q$ составят необходимую базу априорных данных $\{a_m(i)\}$ для вычисления погрешности линейного предсказания согласно выражению (9). Оценка ее дисперсии может быть получена по формуле средней квадратичной величины:

$$\left.\begin{aligned} \sigma_j^2 &= (N-q)^{-1}\sum_{t=q+1}^{N}\varepsilon_j^2(t), \quad j = k-1, \quad k,\\ \sigma_0^2 &= (2N-q)^{-1}\sum_{t=q+1}^{2N}\varepsilon_0^2(t). \end{aligned}\right\} \quad (11)$$

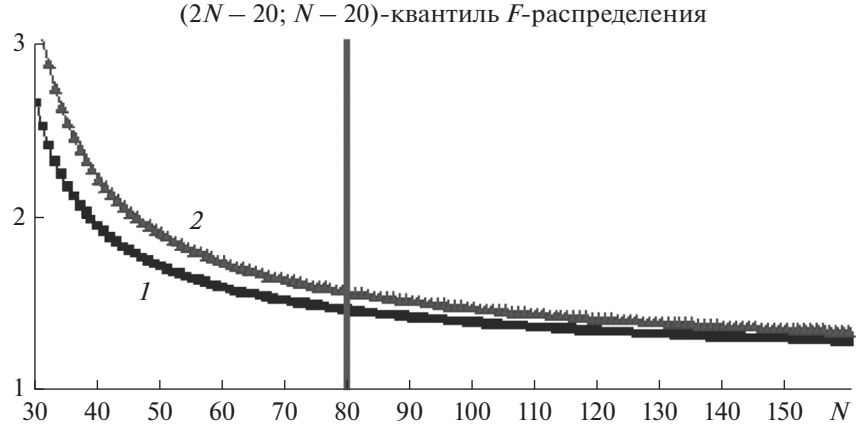

Рис. 1. Зависимости квантиля F-распределения от объема выборки N при различном уровне значимости: $\alpha_0 = 0.1$ (кривая *1*) и 0.05 (кривая *2*).

– в зависимости от анализируемого отрезка речевого сигнала $x(t)$.

Полученный результат (11) совместно с выражениями (8)–(10) определяет адаптивный алгоритм сегментации речевого сигнала $x(t)$ в пределах каждого отдельного речевого фрейма X_k. Его эффективность характеризуется главным образом гарантированным уровнем значимости.

4. АНАЛИЗ ЭФФЕКТИВНОСТИ

Следуя известной методике вычислений [19, 26] и учитывая при этом χ^2-распределение статистик под знаком суммы в правой части (11), а также пренебрежимо малую вероятность случайного события $P\left(A_{k-1} \wedge A_k \middle| H_0\right)$, где $\wedge$ – символ логической конъюнкции, запишем выражение для вероятности ошибки первого рода

$$\alpha_k = P\left(A_{k-1} \vee A_k \middle| H_0\right) \simeq 2P\left(A_j \middle| H_0\right) = \\ = 2P\left(\frac{\sigma_0^2}{\sigma_j^2} > \tilde{\delta}_0 \middle| H_0\right) = 2\left(1 - F_{2N-q;\,N-q}(\tilde{\delta}_0)\right) \qquad (12)$$

при применении критерия (8). Здесь $F_{2N-q;\,N-q}(\tilde{\delta}_0)$ – интегральная функция F-распределения (Фишера) с $(2N-q;\, N-q)$ степенями свободы, $\tilde{\delta}_0$ – установленный наблюдателем пороговый уровень решающей статистики. Указанное распределение подробно табулировано и широко представлено в самых разных источниках, включая электронные таблицы Excel. Например, при $N = 80$, $q = 20$ для уровня значимости $\alpha_0 = 0.10$ (10%) по этим таблицам с помощью функции FРАСПОБР(0.05; $2N - q$; $N - q$) получим порог $\tilde{\delta}_0$, равный квантилю $F_{140;60;0.95} = 1.46$ заданного порядка $1 - \alpha_0/2 = 0.95$. При увеличении объема выборки до $N = 120$ (при длительности речевого фрейма 15 мс) и при сохранении прежнего уровня значимости пороговый уровень может быть понижен до 1.33. Чем меньше величина $\tilde{\delta}_0$, тем строже требования наблюдателя к степени однородности объединенной выборки $\{X_{k-1}; X_k\}$. А установленный при этом уровень значимости α_0 характеризует требования наблюдателя иного рода, а именно: к вероятности ложной отбраковки текущего речевого фрейма X_k как недостаточно четкого, маргинального.

В идеале следует стремиться к минимизации одновременно и $\tilde{\delta}_0$, и α_0. Однако эти требования противоречат друг другу, и поэтому необходимо найти компромисс. Обычно поиск такого компромисса – это самостоятельная задача [7], но только не в нашем случае, когда искомый компромисс очевиден: равенство (12) связывает между собой обе рассматриваемые величины. Отметим при этом монотонность функции распределения $F_{2N-q;\,N-q}(\tilde{\delta}_0)$. Поэтому, устанавливая значение $\tilde{\delta}_0$ в правой части (12) из условия достижения равенства $\alpha_k = \alpha_0$, наблюдатель гарантирует требуемый уровень значимости критерия (8) при минимальном пороговом уровне. О качестве достигаемого в данном случае компромисса свидетельствует рис. 1. На нем представлены два графика зависимости $(2N-q;\ N-q)$-квантиля F-распределения от объема выборки N при заданном порядке $q = 20$ АР-модели речевого сигнала (7) для двух значений уровня значимости α_0 : 5% и 10%. Хорошо видно, что уже при $N = 80$ обе кривые практически утрачивают свою динамику. А это означает, что конечного объема выборки $N = 80$ на интервале наблюдения речевого сигнала в один стандартный фрейм[1] оказывается вполне достаточно для достижения эффекта гарантированного уровня зна-

[1] Стандартный речевой фрейм соответствует стандартной величине (10 мс) периода основного тона.

Рис. 2. Скриншот главного окна компьютерной программы “Phoneme Training”.

чимости без существенных потерь в точности сегментации. Сделанный вывод подтверждается результатами проведенного эксперимента (см. далее).

5. ПРОГРАММА И РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

В подтверждение результатов теоретического анализа был поставлен и проведен эксперимент с использованием авторской компьютерной программы “Phoneme Training”[2]. На рис. 2 показан скриншот ее главного окна. В правой части отображен график оценки спектра мощности гласного звука русской речи “а” методом Берга [23]. При учете его высоких динамических свойств сначала была исследована степень однородности реальных речевых сигналов в обоснование актуальности проведенного выше исследования.

С этой целью в пределах гласных звуков речи достаточно большой длительности (секунды) от контрольного диктора (автора статьи) была сформирована представительная [27] последовательность речевых фреймов длительностью 10 мс каждый. По ним с использованием рекуррентной процедуры (10) были рассчитаны спектральные оценки Берга достаточно большого порядка $q = 20$, которые затем сопоставлялись между собой. Типичные две из них для фонемы “а” представлены на рис. 3. Из их сравнения друг с другом можно сделать вывод о существенной неоднородности речевого сигнала в пределах даже одного звука речи диктора. Это следствие известного [28] эффекта внутридикторской вариативности устной речи. Задача сегментации речевого сигнала приобретает в свете сказанного очевидное практическое значение.

На втором этапе эксперимента программа “Phoneme Training” была переведена для работы в режим “Тестирование”, в котором однофонемные сигналы были подвергнуты автоматической сегментации согласно алгоритму (8)–(11). При этом пороговый уровень разладки $\tilde{\delta}_0$ варьировался в широких пределах с помощью вкладки “Параметры” в меню главного окна (см. рис. 2). Полученные результаты отражены в виде двух временных диаграмм фонемы “а” на рис. 4а, 4б: при $\tilde{\delta}_0 = 1.45$ и $\tilde{\delta}_0 = 1.35$ соответственно. Светлым фоном на рисунке отмечены маргинальные фреймы речевого сигнала, которые не прошли проверку на требуемую степень однородности согласно критерию (8). Их относительная доля в речевом сигнале примерно равна 10% в первом, менее строгом варианте критерия, и 20% – во втором варианте. Оба полученных результата хорошо согласуются с их теоретическими оценками из выражения (12):

[2] См. https://sites.google.com/site/frompldcreators/produkty-1/phonemetraining.

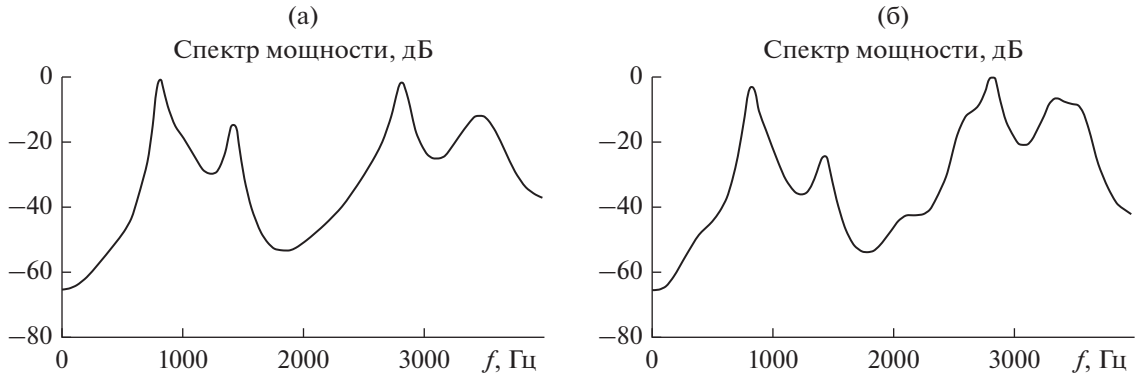

Рис. 3. Скриншот фрагмента главного окна программы “Phoneme Training” с графиком оценки спектра мощности сигнала фонемы “А” в пределах двух разнесенных во времени фреймов.

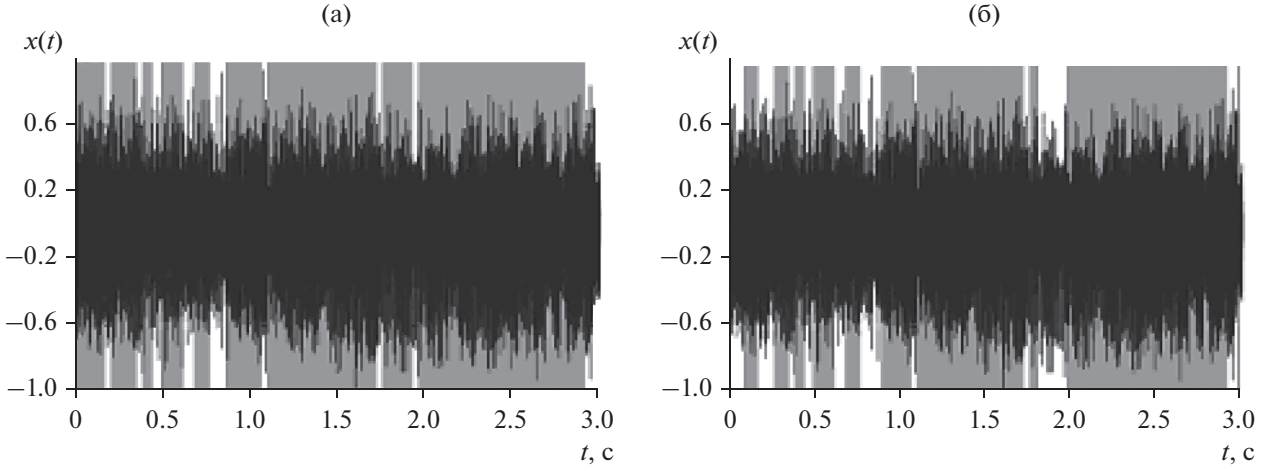

Рис. 4. Скриншот рабочего окна программы в режиме “Тестирование” при двух разных значениях порогового уровня: $\tilde{\delta}_0 = 1.45$ (а) и $\tilde{\delta}_0 = 1.35$ (б).

$\alpha_k = 0.11$ и $\alpha_k = 0.19$ соответственно. При этом статистическая погрешность измерений в ее относительном выражении $\varepsilon = 2/\sqrt{10T/\tau}$ [27] с доверительной вероятностью 0.95 (почти достоверное событие) не превысила в данном случае 3.6%.

6. ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

При анализе письменного текста на русском языке мы опираемся на наши точные знания в отношении количественного и качественного состава используемой фонологической системы, а также закономерностей ее функционирования в разговорной речи. Этими знаниями мы пользуемся, например, при транскрибации потока речи. Однако если мы анализируем звучащий текст на неизвестном языке и нам недоступна полная информация, относящаяся к его тонкой структуре, то можно либо, опираясь на наш лингвистический опыт, давать участкам речевого потока приблизительную интерпретацию в рамках Международного фонетического алфавита[3], либо, обратившись к акустическим понятиям [4], линейно членить речевой сигнал на некие повторяющиеся минимальные единицы и давать им определенные метки. Очевидно, что второй подход со всех точек зрения наиболее информативен и универсален. Именно он и был применен в рамках проведенного выше исследования.

Основная проблема при таком подходе состоит в том [28], что разговорная речь по своим акустическим характеристикам широко варьируется, причем не регулярным образом, не только от одного носителя языка к другому, но и в пределах одного речевого потока от одного диктора. В указанных условиях становится проблематичной сама идея выделения повторяющегося набора фо-

[3] См. http://www.internationalphoneticalphabet.org/.

нетических единиц. Кроме того, их длительность не превышает на практике нескольких десятков миллисекунд, и это главное препятствие для применения традиционных методов теоретической информатики к разговорной речи.

В поисках путей решения перечисленных выше проблем в работах [5, 6] само понятие “фонема” было строго определено в теоретико-информационном смысле как множество однородных фонетических единиц, объединенных в кластер по критерию минимального информационного рассогласования в метрике Кульбака–Лейблера. Условно говоря, человеческий мозг объединяет и запоминает в себе как нечто целое (в виде абстрактного образа) разные образцы (произношения) каждой отдельной фонемы в соответствующей “сфере” своей памяти вокруг абстрактного “центра” с заданным “радиусом”. Этот радиус и определяет в конечном итоге [10] величину порогового уровня $\tilde{\delta}_0$ в предложенном критерии (8). Нетрудно понять, что благодаря такому определению одновременно решается множество проблем в области автоматической обработки речи: и ее вариативности, и априорной неопределенности, и, наконец, проблемы малых выборок наблюдений.

ЗАКЛЮЧЕНИЕ

Таким образом, предложен новый критерий автоматической сегментации речевого сигнала для систем АРР повышенной точности и надежности. Этот критерий без существенных потерь в степени однородности выделяемых сегментов речи гарантирует стабильный уровень значимости при обработке речевых фреймов малой длительности в предположении, что парциальные выборки в совокупности статистически независимы.

Работа выполнена в рамках Программы фундаментальных исследований Национального исследовательского университета “Высшая школа экономики” (НИУ ВШЭ).

СПИСОК ЛИТЕРАТУРЫ

1. *Makhach P., Skarnitzl R.* Principles of Phonetic Segmentation. Praha: Epocha Publ. House, 2013. https://www.researchgate.net/publication/234052076
2. *Pakoci E., Popovic B., Jakovljevic N. et al.* // Lecture Notes in Computer Science. 2016. V. 9811. P. 67.
3. *Попов М.Б.* // Уч. зап. Казан. ун-та. Сер. Гум. науки. 2017. Т. 159. № 5. С. 1144.
4. *Rabiner L.R. Shafer R.W.* Theory and Applications of Digital Speech Processing. Boston: Pearson, 2010.
5. *Савченко В.В.* // РЭ. 2019. Т. 64. № 6. С. 585.
6. *Савченко В.В.* // РЭ. 2018. Т. 63. № 1. С. 60.
7. *Выхованец В.С., Цзяньмин Д.* // Речевые технологии. 2016. № 1. С. 45.
8. *Савченко В.В., Савченко А.В.* Программный комплекс голосового скрытого управления персональным компьютером для дома и офиса. Св-во о государственной регистрации программы для ЭВМ № 2013615628. Опубл. офиц. бюл. “Программы для ЭВМ. Базы данных. Топология интегральных микросхем” № 2 от 20.06.2013.
9. *Савченко А.В., Савченко В.В.* // Измерительная техника. 2019. № 3. Р. 59.
10. *Савченко В.В.* // Изв. вузов. Радиофизика. 2015. Т. 58. № 5. Р. 425.
11. *Савченко В.В.* // Электросвязь. 2017. № 12. С. 22.
12. *Шишкина А.Ф.* // Теория. Практика. Инновации. 2016. № 4. С. 18.
13. *Benati N., Bahi H.* // Proc. 7th Int. Conf. Sci. of Electronics, Technologies of Information and Telecommun (SETIT 2016). Hammamet, 18–20 Dec. N.Y.: IEEE, 2017. P. 267.
14. *Савченко А.В.* // Информ. системы и технологии. 2014. № 2. С. 12.
15. *Sakran A.E., Abdou S.M., Hamid S.E., Rashwan M.* // Int. J. Comp. Sci. Mobile Comp. 2017. V. 6. № 4. P. 308. https://www.researchgate.net/publication/317339722
16. *Kamper H., Jansen A., Goldwater S.* // Computer Speech & Language. 2017. V. 46. P. 154.
17. *Якимук А.Ю. Конев А.А.* // Информатика и системы управления. 2018. № 2. С. 108.
18. *Савченко В.В.* // /Изв. вузов. Радиоэлектроника. 2017. Т. 61. № 9. Р. 536.
19. *Акатьев Д.Ю., Савченко В.В.* // Автометрия. 2005. Т. 41. № 2. С. 68.
20. *Savchenko A.V.* // Lecture Notes in Artificial Intelligence. 2017. V. 10314. P. 264.
21. *Савченко В.В.* // Изв. вузов. Радиофизика. 2017. Т. 60. № 1. Р. 89.
22. *Боровков А.А.* Математическая статистика. СПб.: Лань, 2010.
23. *Марпл С.Л.-мл.* Цифровой спектральный анализ и его приложения. М.: Мир, 1990.
24. *Kullback S.* Information Theory and Statistics. N.Y.: Dover Publ., 1997.
25. *Gray R.M., Buzo A., Gray A.H., Matsuyama Y.* // IEEE Trans. 1980. V. ASSP-28. № 4. P. 367.
26. *Савченко В.В.* // РЭ. 1997. Т. 42. № 4. С. 428.
27. *Савченко В.В.* // Научные ведомости Белгород. гос. ун-та. Серия: История. Политология. Экономика. Информатика. 2015. Т. 33/1. № 1. С. 74. http://dspace.bsu.edu.ru/handle/123456789/12929.
28. *Ронжин А.Л., Евграфова К.В.* // Изв. вузов. Сер. Гум. науки. 2011. Т. 2. № 3. С. 227.