РАДИОТЕХНИКА И ЭЛЕКТРОНИКА, 2023, том 68, № 7, с. 660–668

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ СИГНАЛОВ

УДК 621.391:534.442

СУБОПТИМАЛЬНЫЙ АЛГОРИТМ ИЗМЕРЕНИЯ ЧАСТОТЫ ОСНОВНОГО ТОНА С ИСПОЛЬЗОВАНИЕМ ДИСКРЕТНОГО ФУРЬЕ-ПРЕОБРАЗОВАНИЯ РЕЧЕВОГО СИГНАЛА

© 2023 г. В. В. Савченко[a], *, Л. В. Савченко[b]

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

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

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

Поступила в редакцию 10.03.2022 г.
После доработки 22.12.2022 г.
Принята к публикации 25.02.2023 г.

Отталкиваясь от определения основного тона речи диктора как минимальной частоты линейчатого спектра мощности вокализованных отрезков речевого сигнала, дана оценка потенциально достижимой точности ее измерения в условиях действии фоновых помех типа белого гауссова шума. На основе этой оценки разработан субоптимальный алгоритм измерения частоты основного тона по короткому фрейму речевого сигнала. Эффективность разработанного алгоритма подтверждена результатами проведенного эксперимента, в ходе которого использовалось авторское программное обеспечение.

DOI: 10.31857/S0033849423060128, **EDN:** XMZLPZ

ВВЕДЕНИЕ

Частота основного тона (ЧОТ) относится к числу наиболее информативных акустических характеристик речевого сигнала [1, 2] и в этом качестве широко используется в системах автоматической обработки речи (АОР) различного назначения [3–6]. Как следствие, на протяжении многих лет измерение ЧОТ является классической задачей исследований в области речевых технологий. Неудивительно поэтому, что существует множество разных подходов к ее решению, в рамках которых предложено множество разных алгоритмов [7, 8]. Казалось бы, данное направление исследований должно быть давно исчерпано, по крайней мере в теории. Однако работы, связанные с ЧОТ, имеют явную тенденцию к развитию и даже к расширению в будущем [4–9]. Сказанное объясняется, на наш взгляд, тем, что у разработчиков и пользователей речевых технологий на данный момент отсутствует необходимая ясность в вопросе о потенциально достижимой точности измерения ЧОТ при действии случайных помех в канале связи [10, 11]. Проблема обостряется условиями внутридикторской вариативности речевого сигнала [12–14], а также малых выборок наблюдений на относительно коротких (десятые и сотые доли секунды) отрезках его приблизительной (квази-) стационарности [15–20]. Поэтому актуальной представляется тема исследования, проведенного в рамках данной статьи. Его цель – разработка субоптимального алгоритма измерения ЧОТ на фоне помех типа белого гауссова шума. Для ее достижения используется авторская методика дискретного спектрального моделирования [21].

1. ПОСТАНОВКА ЗАДАЧИ

Согласно акустической теории речеобразования [1] энергетический спектр или спектр мощности $G_x(f)$ вокализованных отрезков (фреймов) речевого сигнала имеет линейчатую структуру [14]. Его минимальная частота F_0, она же – период повторения $K = 0.5F/F_0 \gg 1$ узкополосных (квазигармонических) составляющих речевого сигнала в частотной области (F – частота временной дискретизации), и определяет понятие ЧОТ [3, 4]. Например, в диапазоне значений ЧОТ $F_0 =$ $= 100...200$ Гц дикторов-мужчин [11, 22] при частоте дискретизации речевого сигнала $F = 8$ кГц (соответствует полосе пропускания стандартного телефонного канала связи) получаем $K = 4000/(100...200) =$ $= 20...40$ гармоник с частотами $F_0, 2F_0, ..., KF_0$. Оптимальный алгоритм обработки такого сигнала осуществляется по схеме последовательного соединения двух линейных фильтров [23, 24]: внутрипериодной автокомпенсации речевого сигнала

$x(t)$ и межпериодного накопления сигнала $y(t)$ на выходе автокомпенсатора. В основе их практической реализации используются предикторы речи [24, 25]) двух уровней: в пределах одного периода основного тона $T_0 = 1/F_0 \gg T = 1/F$ и на интервале нескольких таких периодов. Внутрипериодный автокомпенсатор (обеляющий фильтр [17]) нацелен на выравнивание [24]) или обеление [26, 27]) в полосе рабочих частот $f \le 0.5F$ огибающей спектра мощности $G_x(f)$. В идеале на его выходе мы должны получить периодическую (с периодом T_0) последовательность импульсов возбуждения голосового тракта диктора [24, 28]. Для ее выделения из сигнала $y(t)$ и предназначен упомянутый выше межпериодный накопитель. При обработке во временной области он выполняется по схеме рекурсивного (гребенчатого) фильтра. Однако в этом случае остро возникает проблема априорной неопределенности в отношении истинного значения периода основного тона T_0. В теории [29] эту проблему преодолевают путем многоканальной (многоальтернативной) обработки речевого сигнала с настройкой каждого канала на соответствующий возможный вариант периода T_0 [30]. Выбор наилучшей альтернативы осуществляется условным наблюдателем по принципу максимизации мощности сигнала на выходе соответствующего канала. Это весьма затратное во всех отношениях техническое решение, которое плохо сочетается с режимом измерений ЧОТ в реальном времени [9, 31]. Поэтому на практике в роли накопителя импульсов сигнала $y(t)$ широко применяют простейший фильтр нижних частот (ФНЧ) [10, 32] с полосой прозрачности в диапазоне $[\min F_0;\ \max F_0]$ Гц, или близкую к нему в принципиальном отношении кепстральную обработку [25, 33]. В таком случае оконечный блок измерителя выполняется по схеме частотного детектора.

В рамках универсальной стохастической модели речевого сигнала [12] точность измерений ЧОТ может быть охарактеризована дисперсией погрешности σ_F^2. В предположении о стационарности сигнала в пределах речевого фрейма воспользуемся известной формулой [29]

$$\sigma_F^2 = \left(h_z^2 \tau_z^2\right)^{-1}$$

дисперсии погрешности измерений частоты узкополосного (радио) импульса при действии белого гауссова шума. Здесь h_z^2 и τ_z – соответственно отношение сигнал/шум по мощности (ОСШ) и эффективная длительность сигнала $z(t)$ на выходе ФНЧ. Для случая прямоугольной формы его огибающей выполняется равенство $\tau_z = \pi\tau/\sqrt{3}$, где τ – длительность речевого фрейма. При этом в пересчете к среднеквадратичному отклонению (СКО) погрешности измерений получаем

$$\sigma_F = \sqrt{\frac{3}{(\pi\tau)^2 h_z^2}}. \qquad (1)$$

Потребуем от СКО (1) выполнения условия вида $\delta_F = \sigma_F/F_0 \le \delta_0$, где δ_0 – некоторый пороговый (относительный) уровень. В расчете на гауссово распределение погрешности измерений это условие гарантирует с доверительной вероятностью 0.95 длину доверительного интервала на уровне $\Delta_F = 2\sigma_F \le 2\delta_0 F_0$. Например, при $F_0 = 100$ Гц и $\delta_0 = 4...5\%$ получаем $\Delta_F = 8...10$ Гц, что соответствует типовым требованиям пользователей [6–10] к точности измерения ЧОТ в системах АОР. Отсюда вытекает ограничение вида

$$h_z^2 = \frac{3}{(\pi\tau\sigma_F)^2} \ge \left(\frac{2\sqrt{3}}{\pi\tau\delta_0 F_0}\right)^2.$$

Так, в условиях рассматриваемого примера при длительности речевого фрейма $\tau = 20...30$ мс ОСШ h_z^2 на выходе ФНЧ должно быть не ниже порога 18...23 дБ. Отметим, что это практически трудно выполнимое требование, поскольку величина ОСШ определяется в выражении (1) через энергию лишь одной из K гармонических составляющих спектра мощности $G_y(f)$ последовательности $y(t)$ импульсов возбуждения. Соответственно, и энергия сигнала $z(t)$ на выходе ФНЧ в K раз меньше, чем энергия сигнала $y(t)$ на его входе.

Исправить ситуацию кардинальным образом можно за счет суммирования энергии всех $K \gg 1$ гармоник ЧОТ речевого сигнала в спектре мощности $G_y(f)$ отклика обеляющего фильтра $y(t)$. По аналогии с (1) в таком случае будем иметь[1]

$$\sigma_F = \sqrt{\frac{3}{(\pi\tau)^2 K h_z^2}}. \qquad (2)$$

При этом сильно ослабляются требования наблюдателя к интенсивности фонового шума на входе измерителя ЧОТ. Так, при $K = 30$ в условиях предыдущего примера получим выигрыш по ОСШ, превышающий 15 дБ. Однако для его осуществления на практике требуется принципиально иной, по сравнению с межпериодным накоплением, способ обработки: с адаптацией к тонкой структуре речевого сигнала на интервале длительностью $\tau \gg T_0$ в несколько периодов основного тона.

Наиболее естественным способом измерения ЧОТ с суммированием энергии ее гармоник на частотах $F_k = kF_0$, $k \le K$, является обработка речевого сигнала в частотной области – с использованием дискретного фурье-преобразования (ДФП)

[1] На выходе обеляющего фильтра пачка сигнальных составляющих в частотной области имеет прямоугольную форму [9].

[34]. Его размерность $M \gg 1$ определяет избирательную способность ДФП по частоте $\Delta f = F/M$. Поскольку в нашем случае требуется выполнить условие $\Delta f \le \Delta_F$, то при $\Delta_F = (8...10)$ Гц и $F = 8$ кГц получаем $M \ge 1000$. Данному условию отвечает, в частности, алгоритм быстрого преобразования Фурье [29] размерностью $M = 2^{10}$. Проблема вычислительной сложности и быстродействия измерителя ЧОТ при этом практически утрачивает свою актуальность, однако возникает новая проблема – вредное влияние первых формант в спектрах мощности гласных звуков речи [22, 32]. В силу своей относительной интенсивности они подавляют значительную часть гармоник ЧОТ в низкочастотной части спектра $G_y(f)$ и этим сильно искажают результаты измерений [24, 28]. По-видимому, именно данным обстоятельством объясняется общеизвестный феномен [3, 35] так называемых грубых ошибок [1, 24]) в задаче измерения ЧОТ. Между тем существует способ, если не исключить совсем, то сильно ослабить указанную проблему. Он связан с идеей обеления огибающей линейчатого спектра мощности $G_x(f)$ речевого сигнала[2] с использованием авторегрессионной модели (АР-модели) $G_p(f)$ относительно небольшого порядка $p = 8...10$, адаптированной по результатам ДФП на интервале наблюдений длительностью τ. Соответствующая вычислительная процедура, причем в высокоскоростном варианте, подробно описана в работе [21].

2. СИНТЕЗ АЛГОРИТМА

Спектр мощности речевого сигнала, заданного в пределах наблюдаемого фрейма $x(t)$ конечной длительности $\tau = 20...30$ мс последовательностью своих эквидистантных (с периодом $T = 1/F = \text{const}$) отсчетов $x(n) = x(t_n)$ в дискретном времени $t = t_n$, $n = 0, 1, ..., N-1$, где $N = \tau/T$, определяется квадратом модуля спектральной плотности Фурье общего вида:

$$S_x(jf) = T\sum_{n=0}^{N-1} x(n)\exp(-j2\pi nfT), \quad f \le 1/(2T),$$

где j – символ мнимой единицы. При применении M-точечного ДФП выражение для спектра мощности приобретает следующий вид:

$$G_x(f_m) = \left|T\sum_{n=0}^{M-1} x(n)\exp\left(-j2\pi nmM^{-1}\right)\right|^2, \quad m = 0,1,\ldots,M-1, \tag{3}$$

где $f_m = m\Delta f$; $\Delta f = F/M$. Здесь только первые N из $M \gg$ N отсчетов временного ряда $\{x(n)\}$ отличны от нуля [34]. Вместе с тем АР-модель того же сигнала $x(t)$ в частотной области имеет вид инверсного преобразования [18, 19]

$$\begin{aligned}\hat{G}_x(f_m) &= \sigma_x^2 T\left|1-\sum_{i=1}^{p} a_p(i)\exp\left(-j2\pi imM^{-1}\right)\right|^{-2} = \\ &= \sigma_x^2 T\left|\sum_{i=0}^{p} b(i)\exp\left(-j2\pi imM^{-1}\right)\right|^{-2} = \\ &= \sigma_x^2 T\left|B_p(jf_m)\right|^{-2} \triangleq \hat{G}_x(f_m;b_{p+1})\end{aligned} \tag{4}$$

квадрата модуля комплексного коэффициента передачи

$$B_p(jf_m) = \sum_{i=0}^{p} b(i)\exp\left(-j2\pi imM^{-1}\right) \tag{5}$$

линейного трансверсального фильтра p-го порядка. Здесь $a_p(i)$ – i-й элемент вектора $\mathbf{a}_p = \{a_p(i)\}$ коэффициентов линейной авторегрессии того же порядка p; σ_x^2 – масштабный множитель; $b(i)$ – i-й элемент вектора $\mathbf{b}_{p+1} = \{1; -\mathbf{a}_p\}$ весовых коэффициентов. Сигнал $y(t)$ на выходе фильтра (5) в частотной области описывается выражением

$$G_y(f_m) = G_x(f_m)/\hat{G}_x(f_m) = \left|B_p(jf_m)\right|^2 G_x(f_m), \tag{6}$$

в котором множитель σ_x^2 приравнен к единице как не имеющий значения в контексте решаемой задачи. При правильно подобранном в (5) векторе коэффициентов $\mathbf{b}_{p+1} = \{b(i)\}$ и при относительно невысоком порядке p выражение (6) определяет ключевой элемент в составе измерителя ЧОТ: обеляющий фильтр или блок выравнивания огибающей спектра мощности $G_x(f)$. Проблема состоит в том, что при учете внутридикторской вариативности речевого сигнала данный вектор должен быть адаптирован к спектру (3) в режиме скользящего окна наблюдений [12]. В работе [21] в этих целях предложена итеративная (пошаговая) процедура оптимизации градиентного типа

$$\begin{aligned}\mathbf{b}_{p+1}(l) = \mathbf{b}_{p+1}(l-1) - \\ -\gamma_0\nabla\rho_{\text{M-COSH}}(\mathbf{b}_{p+1})\big|_{\mathbf{b}_{p+1}=\mathbf{b}_{p+1}(l-1)}, \\ l = 1,2,\ldots,\end{aligned} \tag{7}$$

с использованием дискретного спектра (3) в качестве эталона. Здесь γ_0 – шаг итераций, l – номер шага; ∇ – знак градиента. В указанной процедуре в роли целевого функционала применяется модификация COSH-расстояния следующего вида:

[2] Понятие огибающей тонкой структуры спектра мощности речевого сигнала, или его спектральной огибающей, широко используется в области АОР и подробно описано, например, в работе [13].

$$\rho_{\text{M-COSH}}(\mathbf{b}_{p+1}) = \\ = \sqrt{\left[M^{-1}\sum_{m=0}^{M-1}\hat{G}_x(f_m;\mathbf{b}_{p+1})G_x^{-1}(f_m)\right]\left[M^{-1}\sum_{m=0}^{M-1}G_x(f_m)\hat{G}_x^{-1}(f_m;\mathbf{b}_{p+1})\right]} - 1 \geq 0. \quad (8)$$

Отличительной особенностью данной модификации является свойство масштабной инвариантности и высокое быстродействие [36]. Как следствие, последовательность приближений (7) сходится в окрестности точки искомого оптимума

$$\mathbf{b}_{\text{opt}} = \text{Arg}\min\rho_{\text{M-COSH}}(\mathbf{b}_{p+1}) \quad (9)$$

всего за $L = 8...16$ итераций [21]. При этом в соответствии с (6) в спектре мощности отклика обеляющего фильтра $y(t)$ будет сформировано $K = F/F_0 \gg 1$ ярко выраженных гармонических составляющих $G_{y,k}(f_m)$ с равномерным сдвигом между собой по частоте – в идеале на F_0. Измеряя величину частотного сдвига тем или иным способом, мы получаем искомую оценку ЧОТ $\hat{F}_0$.

Продуктивным для выявления периодических компонент в сигнале $y(t)$ представляется подход [35, 37], основанный на вычислении автокорреляционной функции (АКФ) спектра (9) на выделенном наборе частот $\{f_m\}$. При условии его предварительного амплитудного квантовании (нормализации) на два уровня:

$$G_z(f_m) = \sum_{k=1}^{K} G_{z,k}(f_m), \\ G_{z,k}(f_m) = \begin{cases} 1, & G_{y,k}(f_m) > g_0, \\ 0, & G_{y,k}(f_m) \leq g_0, \end{cases} \quad (10)$$

где g_0 – некоторый пороговый уровень, получим решающее правило вида

$$\hat{F}_0 = \text{Arg}\max_{r_1 \leq r \leq r_2} Q(\Delta f_r). \quad (11)$$

Здесь $Q(\Delta f_r)$ – нормированная к единице АКФ квантованного спектра мощности (10). Ее аргументом служит частотный сдвиг $\Delta f_r = r\Delta f$ в границах $r_1\Delta f \leq \Delta f_r \leq r_2\Delta f$ анализируемого диапазона частот. Эти границы регулируются выбором двух констант $r_1 \leq r_2 < M$. Причем для исключения тривиального для любой АКФ результата: $\hat{F}_0 = 0$, константа $r_1 \geq 1$ не может быть установлена равной нулю. Напротив, рекомендуется устанавливать соотношение $r_1 = \min F_0/\Delta f \gg 1$. При достижении равенства $f_r = F_0$ зависимость $Q(\Delta f_r)$ реализует эффект суммирования энергии гармоник ЧОТ.

Выражения (3)–(11) в совокупности определяют предлагаемый алгоритм измерения ЧОТ повышенной точности. Вычислитель АКФ играет в нем роль межпериодного накопителя сигнала $y(t)$ в частотной области. Его близость к оптимальному алгоритму по эффективности (2) подтверждается результатами проведенного далее эксперимента.

3. ПРОГРАММА И РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Эксперимент проводился в два этапа. На первом объектом исследования служили сигналы шести гласных звуков русской речи в их аддитивной смеси с белым гауссовым шумом. Дисперсия шума, а вслед за ней и ОСШ h_x^2 на входе измерителя ЧОТ варьировались в эксперименте в пределах 0...40 дБ. Длительность каждого сигнала (2.5...3 с) изначально была рассчитана на его автоматическое членение на множество коротких ($\tau = 30$ мс) фреймов данных $x(t)$ при их частичном (по 20 мс) перекрытии во времени. При этом частота дискретизации F была установлена равной 8 кГц. В результате объем речевой базы данных по каждому гласному звуку речи составил $R = (2.5...3)/10^{-2} = 250...300$ независимых фреймов размерностью $N = 30 \times 8 = 240$ отсчетов речевого сигнала для каждого отдельного значения ОСШ. По каждому фрейму с использованием алгоритма быстрого преобразования Фурье при равенстве длины его окна $M = 2^{10}$ был вычислен с шагом по частоте $\Delta f = 8000/1024 = 7.8125$ Гц текущий (мгновенный) спектр мощности (3) наблюдаемого сигнала $x(t)$. Затем согласно (4) была построена АР-модель его огибающей. По ней в соответствии с (10) был сформирован линейчатый спектр $G_y(f)$ последовательности $y(t)$ импульсов возбуждения голосового тракта контрольного диктора. После этого путем автокорреляционной обработки (11) была получена выборочная оценка ЧОТ. Ее инструментальная погрешность измерений $0.5\Delta f$ была сокращена до 1...2 Гц путем интерполяции зависимости $Q(\Delta f_r)$ в окрестностях локального максимума. В дальнейшем она была подвергнута совместно с аналогичными оценками $\hat{F}_0$ в пределах R-выборки экспериментальных данных $\{x(t)\}$ статистическому усреднению. В итоге при доверительной вероятности 0.9 погрешность выборочной оценки ЧОТ в ее относительном выражении $\varepsilon_F = 1.65/\sqrt{R} = 1.65/\sqrt{(250...300)}$ составила 9...10% [17, 38].

Все основные вычисления в ходе эксперимента были выполнены с использованием новейшей модификации авторской компьютерной программы Phoneme Training[3]. Полученные результаты отражены в виде графиков и диаграмм на рис. 1–6.

[3] Программа размещена в режиме открытого доступа на сайте авторов статьи по ссылке https://sites.google.com/site/frompldcreators/produkty-1/phonemetraining.

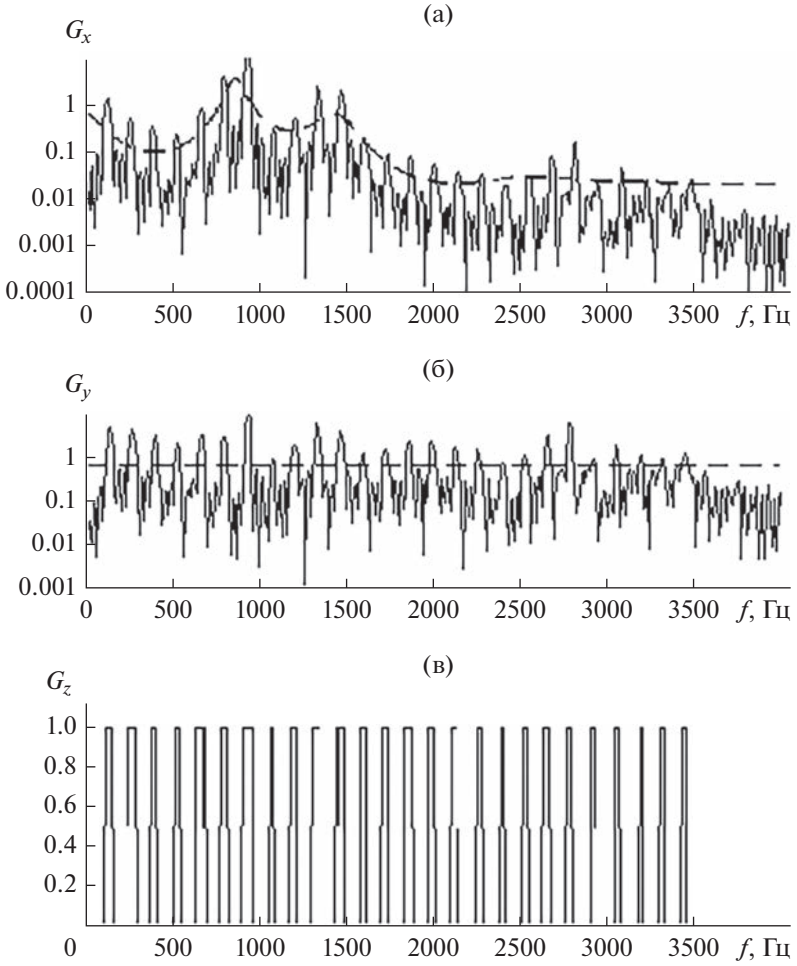

Рис. 1. Спектр мощности сигнала гласного звука “а” (сплошная линия) и его огибающая (штриховая линия) на входе измерителя ЧОТ (а), на выходе обеляющего фильтра (б) и выходе амплитудного квантователя (в).

На рис. 1а–1в представлено типичное семейство спектров мощности гласного звука “а” от контрольного диктора в условиях относительно небольшого ($h_x^2 = 30$ дБ) фонового шума – на выходах спектрального анализатора (3), обеляющего фильтра (6) и амплитудного квантователя (10) соответственно. Штриховой линией на рисунке показана форма огибающей спектра мощности речевого сигнала. Из сравнения графиков на рис. 1а и 1б друг с другом хорошо видны эффекты как спектрального АР-моделирования, так и выравнивания спектральной огибающей [9]. Горизонтальной линией на втором графике отмечен пороговый уровень g_0, равный в данном случае средней мощности сигнала $y(t)$ в частотной области:

$$P_y = M^{-1}\sum_{m=0}^{M-1} G_y(f_m).$$

При этом график на рис. 1в наглядно отражает назначение амплитудного квантователя (10) в составе субоптимального измерителя ЧОТ, а именно: нормализация по форме и по уровню гармонических составляющих спектра (9). А достигаемый эффект в полной мере отражают графики нормированной АКФ спектра мощности речевого сигнала на рис. 2а и 2б, построенные согласно (11) при равенстве $h_x^2 = 30$ и 10 дБ соответственно. Как видим, амплитуда

$$q_F = Q(\Delta f_r)\big|_{\Delta f_r = \hat{F}_0} \qquad (12)$$

первого пика АКФ в области боковых лепестков (на рис. 2 отмечен темным квадратом) существенным образом зависит от ОСШ. Однако его положение в частотной области, а, значит, и результат $\hat{F}_0 \approx 132.8$ Гц измерений ЧОТ, стабильны в обоих случаях. Это признак достаточно высокой помехоустойчивости предложенного измерителя.

В развитие данного вывода на рис. 3 представлен график зависимости СКО оценки ЧОТ (11) от ОСШ при действии гласного звука речи “а” на фоне белого шума (кривая *1*). Для сравнения здесь же представлен график потенциально до-

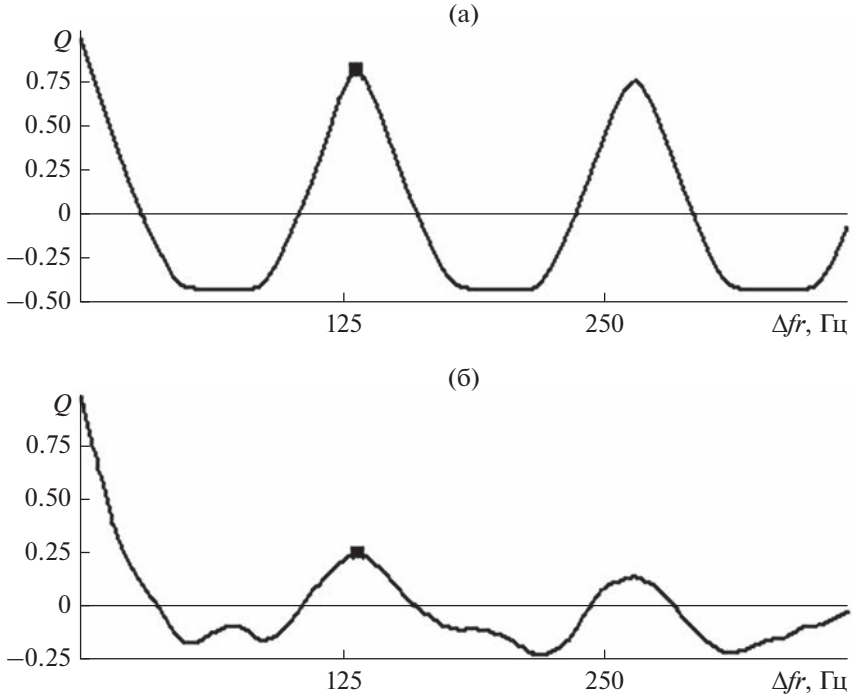

Рис. 2. Нормированная АКФ спектра мощности речевого сигнала на выходе амплитудного квантователя (6) при ОСШ, равном 30 (а) и 10 дБ (б).

стижимой СКО (2) при равенстве $K = 30$ (кривая *2*). При этом ОСШ h_z^2 было пересчитано в ОСШ на входе измерителя по формуле $h_x^2 = h_x^2 + 10$, дБ. В ней учтен известный эффект ослабления речевого сигнала на выходе автокомпенсатора [18]. Его количественная характеристика – в данном случае 10 дБ – получена экспериментальным путем [19] применительно к фонеме “а” от контрольного диктора. Как видим, оба графика мало отличаются друг от друга в области рабочих значений ОСШ $h_x^2 \geq 10$ дБ. Напротив, они сильно разнятся с графиком аналогичной зависимости (1) в отсутствие эффекта межпериодного накопления (кривая *3*). По сравнению с ним субоптимальный алгоритм характеризуется выигрышем в пороговых сигналах порядка 8...12 дБ.

Сделанные выводы сохраняют справедливость для всех остальных, наряду с фонемой “а”, гласных звуков речи контрольного диктора. Правда, оценки ЧОТ при этом могут сильно варьироваться по величине. Например, для звука “и” от контрольного диктора результат $\hat{F}_0$ составил 115...118 Гц. С указанной точки зрения не только теоретическое, но и практическое значение имеют результаты второго, заключительного этапа эксперимента.

В качестве его объекта исследования была взята строка “Ночь, улица, фонарь, аптека” (А. Блок) в произнесении контрольного диктора. Аудиозапись, сделанная в закрытом помещении при ОСШ порядка 30 дБ, была разделена на последовательность фреймов $x(t)$ длительностью по $\tau = 30$ мс. По каждому из них была осуществлена обработка сигнала согласно алгоритму (3)–(10). Полученные результаты представлены в виде семейства временных диаграмм на рис. 4. Синхронно с речевым сигналом (рис. 4а) показана (рис. 4б) последовательность формируемых согласно (11)

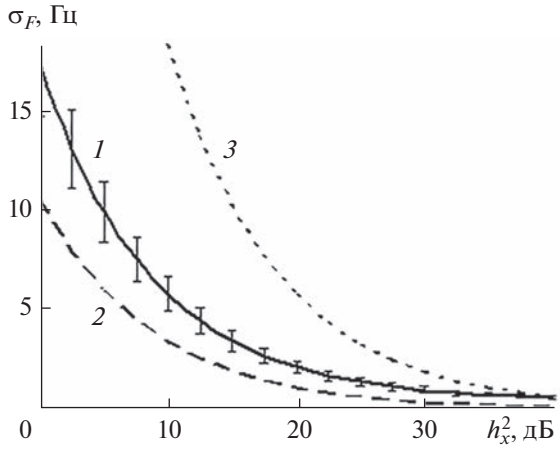

Рис. 3. Зависимость СКО погрешности измерения ЧОТ от ОСШ: кривая *1* – по результатам проведенного эксперимента; кривая *2* – в потенциально достижимом варианте (2) при равенстве $K = 30$; кривая *3* – в отсутствие эффекта межпериодного накопления (1).

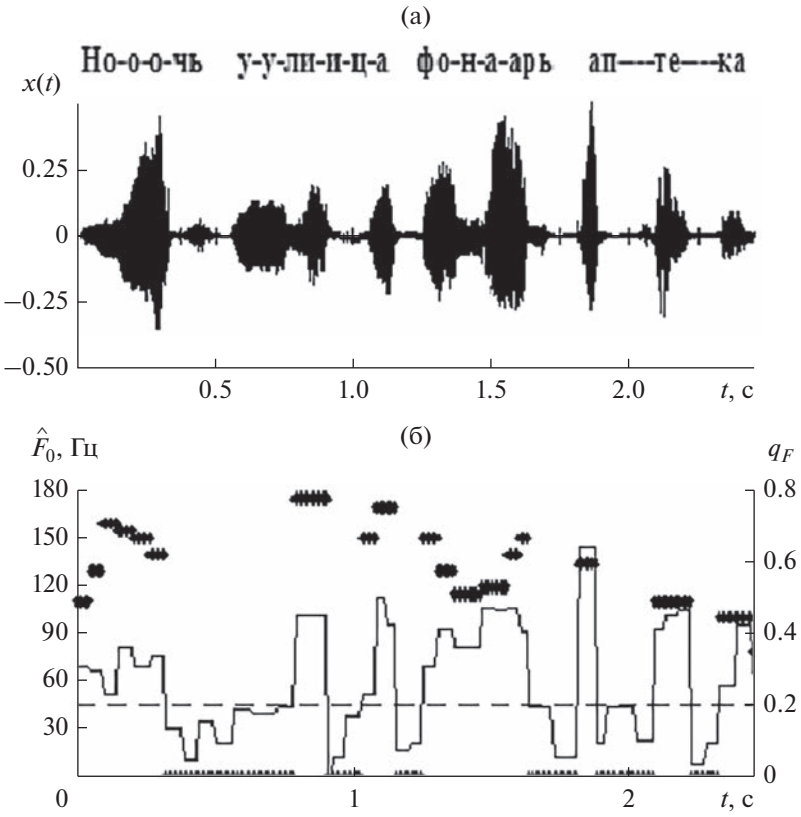

Рис. 4. Временная диаграмма контрольного речевого сигнала (а) и траектория ЧОТ в динамике (б).

оценок ЧОТ (отмечены маркером), а также соответствующая временная диаграмма амплитуды первого пика АКФ. Чем больше эта амплитуда q_F, тем выше надежность текущего результата $\hat{F}_0$.

Как видим, оценки ЧОТ сильно разнятся между собой по надежности. При этом разрывы и пропуски на диаграмме оценок ЧОТ (рис. 4б) указывают на соответствующие недостаточно вокализованные, в частности назальные звуки речи диктора. Причем величина ЧОТ в динамике зависит не только от типа фонемы, но и от интонации диктора в момент ее произнесения.

4. ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Измерение (отслеживание) ЧОТ речевого сигнала является нетривиальной задачей, поскольку до сих пор существует ряд нерешенных проблем. Так, например, в силу своего определения частоту F_0 можно легко перепутать с ее первой гармоникой (эффект удвоения тона “pitch doubling” [24]) или с первой формантой в спектре гласного звука речи. Кроме того, измерить ЧОТ с высокой точностью сложно по аудиозаписям невысокого качества, так как нужный пик на низких частотах спектра мощности в этом случае сильно ослабевает [7, 8]. Предложенный в данной статье алгоритм решает обе проблемы принципиальным образом: в нем перед корреляционной обработкой [35] гармоники ЧОТ выравниваются друг с другом по амплитуде путем адаптации их огибающей (4) под огибающую спектра (3). Достаточно высокая скорость сходимости итеративной процедуры (7), лежащей в основе указанной адаптации, обусловлена использованием масштабно-инвариантной меры (8) информационного рассогласования в качестве целевого функционала оптимизационной задачи [21].

Еще одно положительное качество предложенного алгоритма состоит в возможности автоматической регистрации вокализованных фреймов речевого сигнала по признаку превышения некоторого порогового уровня q_0 боковыми пиками АКФ из выражения (12). В нашем примере (см. рис. 2) этот порог равен $q_0 = 0.2...0.25$, и по нему нетрудно отделить вокализованные фреймы речи диктора от глухих или недостаточно вокализованных фреймов. Напротив, для большинства известных алгоритмов измерений ЧОТ процедура классификации такого рода является одной из основных составляющих их вычислительной сложности [22, 33].

ЗАКЛЮЧЕНИЕ

В результате проведенного исследования разработан субоптимальный алгоритм измерения ЧОТ речевого сигнала, в котором реализован эффект суммирования энергии гармоник. Его ключевое звено – обеляющий фильтр спектральной огибающей – формируется на основе ДФП методом дискретного спектрального моделирования. Эффективность предложенного алгоритма характеризуется высокой степенью близости его характеристик к потенциально достижимым характеристикам эффективности.

Полученные результаты предназначены для применения в системах АОР различного назначения, включая речевые коммуникации [13, 14], прикладную акустику [7, 8] и другие, в целях отслеживания в динамике изменений эмоционального состояния их пользователей по колебаниям ЧОТ в режиме реального времени [38, 39].

ФИНАНСИРОВАНИЕ РАБОТЫ

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, 2011.
2. *Hirst D., Looze C.* // Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge Univ. Press. 2021. P. 336. https://doi.org/10.1017/9781108644198.014
3. *Schenkman B.N., Gidla V.K.* // Appl. Acoustics. 2020. V. 163. Article 107214. https://doi.org/10.1016/j.apacoust.2020.107214
4. *Allam A.R., Ashour A.S., Elnaby M.A., El-Samie F.E.* // 7th Int. Japan-Africa Conf. Electronics, Communications and Computations (JAC-ECC). 2019. P. 106. https://doi.org/10.1109/JAC-ECC48896.2019.9051338
5. *Souza G.V., Duarte J.M., Viegas F. et al.* // J. Voice. 2020. V. 34. № 4. P. 641. https://doi.org/10.1016/j.jvoice.2018.12.007
6. *Stahl J., Mowlaee P.* // Speech Communication. 2019. V. 111. P. 1. https://doi.org/10.1016/j.specom.2019.05.001
7. *Sharma G., Umapathy K., Krishnan S.* // Appl. Acoustics. 2020. V. 158. Article No 107020. https://doi.org/10.1016/j.apacoust.2019.107020
8. *Zhang W., Wang R., Zhang Q., Fang S.* // Appl. Acoustics. 2020. V. 166. Article No 107338. https://doi.org/10.1016/j.apacoust.2020.107338
9. *Савченко А.В., Савченко В.В.* // Измерит. техника. 2022. № 6. С. 60. https://doi.org/10.32446/0368-1025it.2022-6-60-66
10. *Yadav I.C., Shahnawazuddin S., Pradhan G.* // Digital Signal Processing. 2019. V. 86. P. 55. https://doi.org/10.1016/j.dsp.2018.12.013
11. *Kumar S.* // Int. J. Speech Technol. 2019. V. 22. P. 885. https://doi.org/10.1007/s10772-019-09634-5
12. *Savchenko V.V.* // Radioelectronics and Communications Systems. 2020. V. 63. P. 532. https://doi.org/10.3103/S0735272720100039
13. *Tohyama M.* // Acoustic Signals and Hearing. Kanagawa, Japan: Acad. Press, 2020. P. 89. https://doi.org/10.1016/B978-0-12-816391-7.00013-9
14. *Gibson J.D.* // Information. 2016. V. 32. № 7. https://doi.org/10.3390/info7020032
15. *Gu Yu., Wei H.L.* // Inform. Sci. 2018. V. 451–452. P. 195. https://doi.org/10.1016/j.ins.2018.04.007
16. *Cui S., Li E., Kang X.* // IEEE Int. Conf. Multimedia and Expo (ICME). London: United Kingdom. 2020. P. 1. https://doi.org/10.1109/ICME46284.2020.9102765
17. *Smith S.R.* // J. Acoustical Soc. Amer. 2021. V. 150. Article No. A113. https://doi.org/10|1121|10|0007806|
18. *Савченко В.В., Савченко А.В.* // РЭ. 2020. Т. 65. № 11. С. 1101. https://doi.org/10.31857/S0033849420110157
19. *Savchenko V.V., Savchenko A.V.* // Radioelectronics and Commun. Systems. 2019. V. 62. № 5. P. 276. https://doi.org/10.3103/S0735272719050042
20. *Kashani H.B., Sayadiyan A.* // Computer Speech & Language. 2018. V. 50. P. 105. https://doi.org/10.1016/j.csl.2017.12.008
21. *Савченко В.В., Савченко Л.В.* // РЭ. 2021. Т. 66. № 11. С. 1100. https://doi.org/10.31857/S0033849421110085
22. *Kent R.D., Vorperian H.K.* // J. Commun. Disorders. 2018. V. 74. P. 74. https://doi.org/10.1016/j.jcomdis.2018.05.004
23. *Gibson J.D.* // Information. 2019. V. 179. № 10. https://doi.org/10.3390/info10050179
24. Markel J.D., Gray A.H. // Linear Prediction of Speech. Communication and Cybernetics. Berlin; Springer, 1976. V. 12. https://doi.org/10.1007/978-3-642-66286-7_8
25. *Sueur J.* // Sound Analysis and Synthesis with R. Cham: Springer, 2018. https://doi.org/10.1007/978-3-319-77647-7_12
26. *Esfandiari M., Vorobyov S.A., Karimi M.* // Signal Processing. 2020. V. 171. Article No 107480. https://doi.org/10.1016/j.sigpro.2020.107480
27. *Jaramillo A.E., Nielsen J.K., Christensen M.G.* // 27th Europ. Signal Processing Conf. (EUSIPCO). 2019. P. 1. https://doi.org/10.23919/EUSIPCO.2019.8902763
28. *Palaparthi A., Titze I.R.* // Speech Communication. 2020. V. 123. P. 98. https://doi.org/10.1016/j.specom.2020.07.003
29. Радиоэлектронные системы. Основы построения и теория: Справочник / Под ред. Я.Д. Ширмана. 2-е изд. М.: Радиотехника, 2007.
30. *Sinha R., Shahnawazuddin S.* // Computer Speech & Language. 2018. V. 48. P. 103. https://doi.org/10.1016/j.csl.2017.10.007
31. *Zeremdini J., Messaoud M., Bouzid A.* // Appl. Acoustics. 2017. V. 120. P. 45. https://doi.org/10.1016/j.apacoust.2017.01.013

32. *Jouvet D., Laprie Y.* // 25th Eur. Signal Processing Conf. (EUSIPCO). 2017. P. 1614. https://doi.org/10.23919/EUSIPCO.2017.8081482

33. *Oppenheim A.V., Schafer R.W.* // IEEE Signal Processing Magazine. 2004. V. 21. № 5. P. 95. https://doi.org/10.1109/MSP.2004.1328092

34. *Marple S.L.* Digital spectral analysis with applications. 2-nd ed. Mineola, N.Y.: Dover Publications, 2019.

35. *Parlak C., Altun Yu.* // Mathematical Problems in Engineering. 2021. V. 2021. Article No. 6658951. https://doi.org/10.1155/2021/6658951

36. *Savchenko A.V., Savchenko V.V. & Savchenko L.V.* // Optimization Lett. 2021. № 7. P. 1. https://doi.org/10.1007/s11590-021-01790-5

37. *Levkov D.G., Panin A.G., Tkachev I.I.* // The Astrophysical J. 2022. V. 925. №. 2. P. 109. https://doi.org/10.3847/1538-4357/ac3250

38. *Савченко А.В., Савченко Л.В.* // Измерит. техника. 2021. № 4. С. 72. https://doi.org/10.32446/0368-1025it.2021-4-49-57

39. *Akçay M.B., Oğuz K.* // Speech Communication. 2020. V. 116. P. 56. https://doi.org/10.1016/j.specom.2019.12.001