© 2019 г. А.С. КОЛОКОЛОВ, канд. техн. наук (kolokolo@ipu.ru),
И.А. ЛЮБИНСКИЙ, канд. техн. наук (liubinsk@ipu.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

ИЗМЕРЕНИЕ ОСНОВНОГО ТОНА РЕЧЕВОГО СИГНАЛА С ИСПОЛЬЗОВАНИЕМ ФУНКЦИИ АВТОКОРРЕЛЯЦИИ

Предложен метод измерения основного тона речевого сигнала, основанный на получении и последующей обработке автокорреляционной функции анализируемого сигнала, подчеркивающей ее пик, связанный с периодом сигнала. Используемая обработка предотвращает грубые ошибки при измерении основного тона и представляет собой разновидность клиппирования положительных пиков автокорреляционной функции.

Ключевые слова: речевой сигнал, обработка и анализ сигналов.

DOI: 10.1134/S0005231019020090

1. Введение

Согласно теории речеобразования [1] вокализованный речевой сигнал образуется в результате прохождения через речевой тракт импульсов давления, создаваемых голосовыми связками гортани. Импульсы $e(t)$, генерируемые связками, имеют форму, близкую к треугольной, и образуют последовательность импульсов, следующих с интервалом $T_0 = 1/f_0$, где f_0 — основной тон. Таким образом, образование вокализованного речевого сигнала $s(t)$ можно представить сверткой $s(t) = e(t) * h(t)$, где $*$ — операция свертки, $h(t)$ — импульсная характеристика речевого тракта. Последний факт поясняет рис. 1.

Более сложный вид $s(t)$ по сравнению с $e(t)$ обусловлен наличием ряда резонансов (формант) в передаточной функции речевого тракта. Обычно в речевом сигнале наиболее выраженными являются две первые форманты с частотами F_1 и F_2. Из рис. 1 можно видеть, что периодичность, наиболее отчетливо выраженная в $e(t)$, сохраняется в $s(t)$, хотя ее измерение осложняется наличием дополнительных затухающих колебаний вследствие фильтрации $e(t)$ речевым трактом.

В восприятии речевого сигнала частоты f_0, F_1 и F_2 играют разную роль. Частотами формант F_1 и F_2, отражающими изменения геометрии речевого тракта, обычно переносится фонетическая информация в речевом сообщении, в то время как с помощью частоты основного тона f_0 передается информация о характере высказывания, о положении смысловых групп в речевом потоке, об эмоциональном и психофизиологическом состоянии говорящего. Однако в тональных языках информация о гласном может кодироваться также абсолютным значением частоты f_0. Это свидетельствует о важности измерения основного тона в процессе анализа и распознавания речевого сигнала.

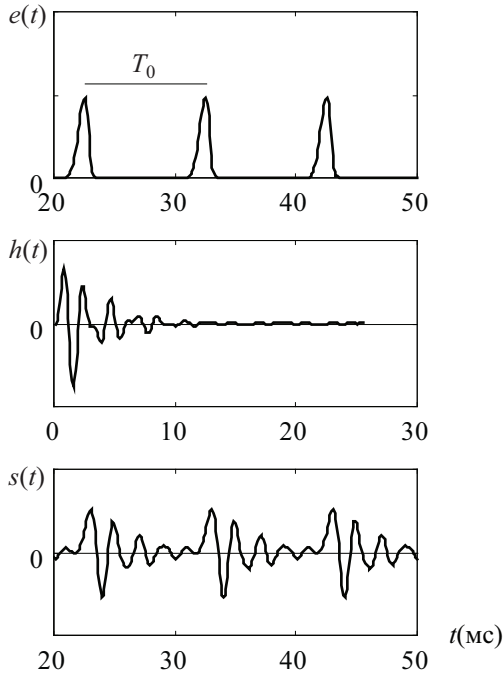

Рис. 1. Иллюстрация формирования речевого сигнала во временной области.

В литературе описывается большое разнообразие методов измерения частоты основного тона, позволяющих производить оценку частоты основного тона во временной, частотной и частотно-временной областях [2–8].

Методы, реализующие измерение основного тона во временной области, базируются как на анализе интервалов между пиками тонкой временной структуры речевой волны, так и на анализе пиков ее автокорреляционной функции, исходного речевого сигнала или речевого сигнала, подвергнутого специальной обработке, усиливающей его периодическую составляющую, связанную с частотой основного тона [9–12].

Оценка частоты основного тона с помощью частотных методов производится на основе измерения частот гармоник речевого сигнала, выявленных с помощью дискретного преобразования Фурье или вейвлет анализа [7, 8].

Примером частотно-временных методов являются методы, основанные на использовании кепстра, представляющего собой косинус-преобразование Фурье от логарифма амплитудного спектра выборки речевого сигнала [4, 5].

Наконец также следует отметить, что одновременно с упомянутыми выше методами также развиваются методы измерения основного тона, базирующиеся на моделях восприятия звука, разрабатываемых с использованием знаний о процессах обработки звукового сигнала в слуховом анализаторе [6].

В рамках настоящей работы ограничимся рассмотрением автокорреляционных измерителей частоты основного тона, основанных на анализе пиков автокорреляционной функции вокализованных фрагментов речевого сигна-

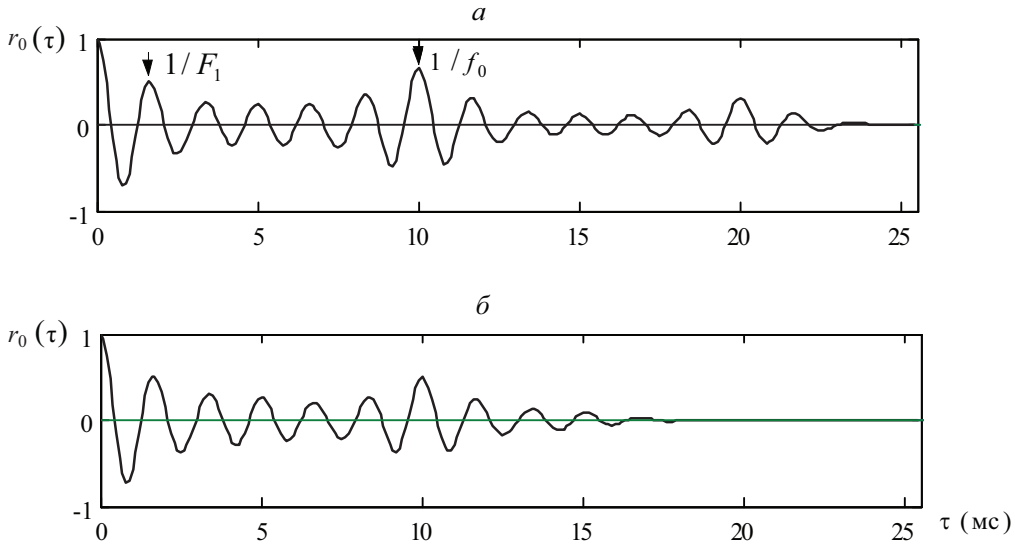

Рис. 2. Вид автокорреляционной функции речевого сигнала $r_0(\tau)$ для случаев: *а* — $\Delta T = 25{,}6$ мс, *б* — $\Delta T = 18$ мс.

ла фиксированной длительности $\Delta T = 20$–50 мс [9, 10], на которых речевой сигнал может рассматриваться как квазистационарный. В этом случае частота основного тона $f_0 = 1/T_0$ для каждого сегмента обычно определяется величиной, обратной координате главного пика на $\tau = T_0$ автокорреляционной функции

$$r(\tau) = \frac{1}{\Delta T} \int\limits_0^{\Delta T - \tau} s(t)s(t-\tau)dt$$

или нормированной функции автокорреляции $r_0(\tau) = r(\tau)/r(0)$. Вид автокорреляционной функции, полученной для гласного, демонстрирует рис. 2. Приведенная на рис. 2,*а* автокорреляционная функция имеет главный пик на $\tau = T_0$, определив положение которого можно найти частоту основного тона $f_0 = 1/T_0$. Однако вследствие того, что речевой сигнал является сверткой сигнала голосового источника с импульсной характеристикой речевого тракта, а автокорреляционная функция для периодического сигнала убывает с τ, главным пиком в отдельных случаях может оказаться пик автокорреляционной функции, связанный с первой формантой F_1 речевого сигнала. Это будет приводить к нежелательным грубым ошибкам измерения основного тона. В частности, такие ошибки могут иметь место при низких значениях f_0, когда $T_0 < \Delta T < 2T_0$. В этом случае амплитуда пика автокорреляционной функции $r(\tau)$ при $\tau = T_0$ может быть меньше амплитуды пика при $\tau = 1/F_1$. Рассмотренную ситуацию поясняет рис. 2,*б*.

Для уменьшения амплитуды пика $r(\tau)$, связанного с первой формантой F_1, может быть использовано центральное клиппирование речевого сигнала [10, 11]. После центрального клиппирования $s(t)$ получается клиппиро-

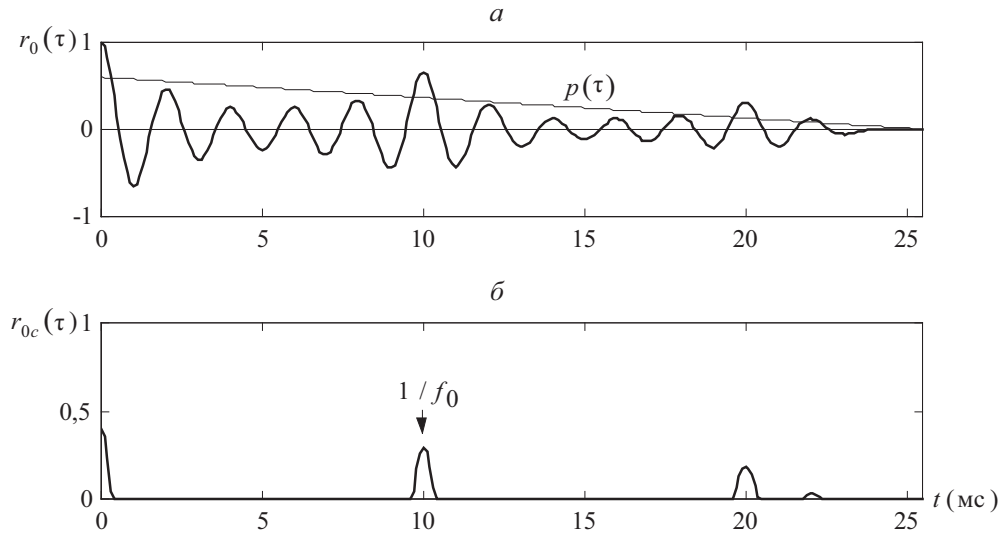

Рис. 3. Клиппирование автокорреляционной функции с помощью линейно-убывающей функции.

ванный сигнал

$$s_c(t) = \begin{cases} s(t) - c_0 & \text{при} \quad s(t) - c_0 \geq 0, \\ s(t) + c_0 & \text{при} \quad s(t) + c_0 \leq 0, \\ s(t) = 0 & \text{при} \quad s(t) - c_0 < 0, \\ s(t) = 0 & \text{при} \quad s(t) + c_0 > 0, \end{cases}$$

где c_0 — уровень клиппирования. Благодаря применению клиппирования, в сигнале $s_c(t)$ и его автокорреляционной функции выравниваются амплитуды гармоник сигнала и тем самым ослабляются его формантные резонансы. В результате подчеркиваются пики в автокорреляционной функции сигнала $s_c(t)$ на $\tau = T_0$ для стационарных участков речевого сигнала. Однако рассмотренная процедура корреляционного анализа с применением клиппирования оказывается неудовлетворительной при изменении амплитуды речевого сигнала на протяжении интервала ΔT и в присутствии импульсных помех.

Ослабить пик в автокорреляционной функции, связанный с первой формантой F_1, можно также с помощью клиппирования положительных пиков в самой автокорреляционной функции [12]. Последнее достигается за счет использования линейно убывающей пороговой функции $p_0(\tau) = \alpha \frac{1}{\Delta T}(\tau - \Delta T)$, где α — параметр, определяющий уровень клиппирования $r_0(\tau)$, выбираемый в диапазоне $0 < \alpha < 1$, а $\tau \in [0,\, \Delta\, T]$. В результате получается клиппированная автокорреляционная функция

$$r_{0c}(\tau) = \begin{cases} r_0(\tau) - \alpha p_0(\tau) & \text{при} \quad r_0(\tau) - \alpha p_0(\tau) > 0, \\ 0 & \text{при} \quad r_0(\tau) - \alpha p_0(\tau) \leq 0. \end{cases}$$

Клиппирование автокорреляционной функции $r_0(\tau)$ с помощью линейно-убывающей функции $p_0(\tau)$ поясняет рис. 3.

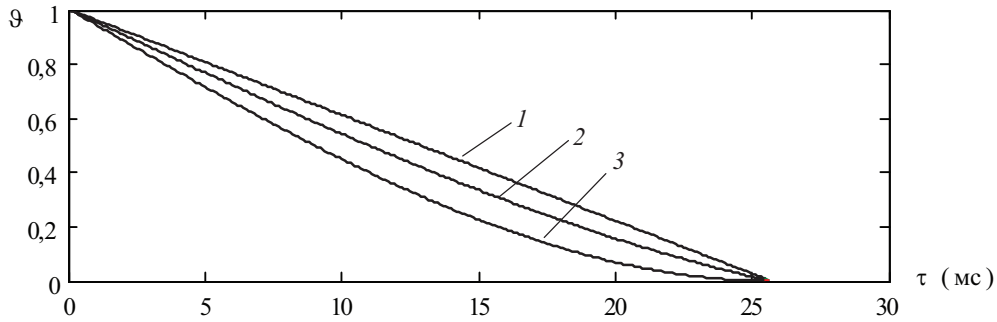

Рис. 4. Влияние изменения амплитуды гармонического сигнала на огибающую его автокорреляционной функции. *1* — $a = 1$, *2* — $a = 0{,}25$, *3* — $a = 0$.

Рассмотренная процедура клиппирования подчеркивает пик автокорреляционной функции на $\tau = T_0$ для стационарных участков речевого сигнала и является малочувствительной к присутствию импульсных помех. Однако она оказывается неудовлетворительной при изменении амплитуды речевого сигнала на протяжении интервала ΔT, так как в этом случае автокорреляционная функция $r_0(\tau)$ может убывать быстрее, чем линейная пороговая функция $p_0(\tau)$. Последнее приводит к пропаданию пика на $\tau = T_0$ и, как следствие, вообще к потере информации о f_0.

Для демонстрации сказанного на рис. 4 приведены примеры влияния изменения амплитуды гармонического сигнала на интервале ΔT на характер убывания огибающей его автокорреляционной функции. Для этого были найдены огибающие автокорреляционные функции гармонического сигнала с постоянной амплитудой и с линейно убывающей амплитудой. Убывание амплитуды создавалось умножением выборки гармонического сигнала длительностью ΔT на окно

$$w(t) = \frac{a-1}{\Delta T}t + 1,$$

где параметр $0 \le a < 1$ определяет конечную амплитуду гармонического сигнала на интервале ΔT. Из рис. 4 можно видеть, что в случае гармонического сигнала с постоянной амплитудой ($a = 1$) его автокорреляционная функция убывает линейно, в то время как отклонение от линейного убывания тем больше, чем быстрее убывает амплитуда гармонического сигнала ($a = 0{,}25$ и $a = 0$) на интервале ΔT.

Ниже предлагается способ повышения надежности измерения частоты основного тона f_0 речевого сигнала путем применения дополнительной обработки автокорреляционной функции $r_0(\tau)$, подчеркивающей ее пик на $\tau = T_0$. Новизна предлагаемого способа подтверждена патентом [13].

2. Описание способа

Суть предлагаемого способа состоит в том, что производится обработка автокорреляционной функции с помощью вычитания из автокорреляционной функции $r_0(\tau)$, полученной для сегмента сигнала, меньшей по амплитуде

сглаженной функции автокорреляции для модуля сигнала на том же сегменте и обнуления отрицательных разностей.

В результате обработки $r_0(\tau)$ получается модифицированная автокорреляционная функция

$$r_{c1}(\tau) = \begin{cases} r_0(\tau) - \alpha r_{0e}(\tau) * h(\tau) & \text{при} \quad r_0(\tau) - \alpha r_{0e}(\tau) * h(\tau) > 0, \\ 0 & \text{при} \quad r_0(\tau) - \alpha r_{0e}(\tau) * h(\tau) \le 0, \end{cases}$$

где $r_{0e}(\tau) = \int_0^{\Delta T - \tau} |s(t)|\,|s(t-\tau)|\,dt / \int_0^{\Delta T} s^2(t)dt$; $h(\tau)$ — симметричная импульсная характеристика сглаживающего фильтра, которая в частном случае отсутствия сглаживания будет представлять собой δ-функцию Дирака; $0 < \alpha < 1$; $\tau \in [0,\, \Delta T]$; $|s(t)|$ — модуль $s(t)$.

Такого рода обработку можно рассматривать как разновидность клиппирования $r_0(\tau)$ с пороговой функцией $\alpha r_{0e}(\tau)$, затухающей примерно так же, как и $r_0(\tau)$. В результате этого у функции $r_{c1}(\tau)$ амплитуда пика на $\tau = T_0$ меньше зависит от изменения амплитуды речевого сигнала на протяжении интервала ΔT, чем у функций $r_0(\tau)$ и $r_{0c}(\tau)$. Благодаря этому, снижается количество ошибок при измерении основного тона.

3. Результаты исследования

Предложенный способ был проверен на фрагментах синтетических и естественных гласных, взятых из речевого сигнала. Сравнительный анализ каче-

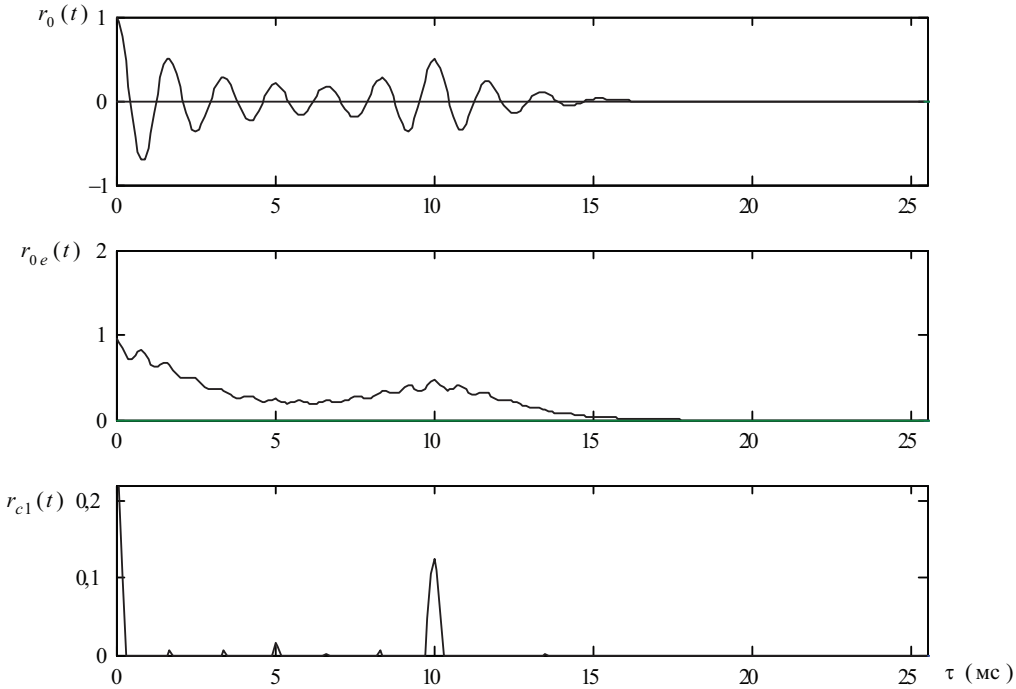

Рис. 5. Иллюстрация предлагаемого способа обработки автокорреляционной функции.

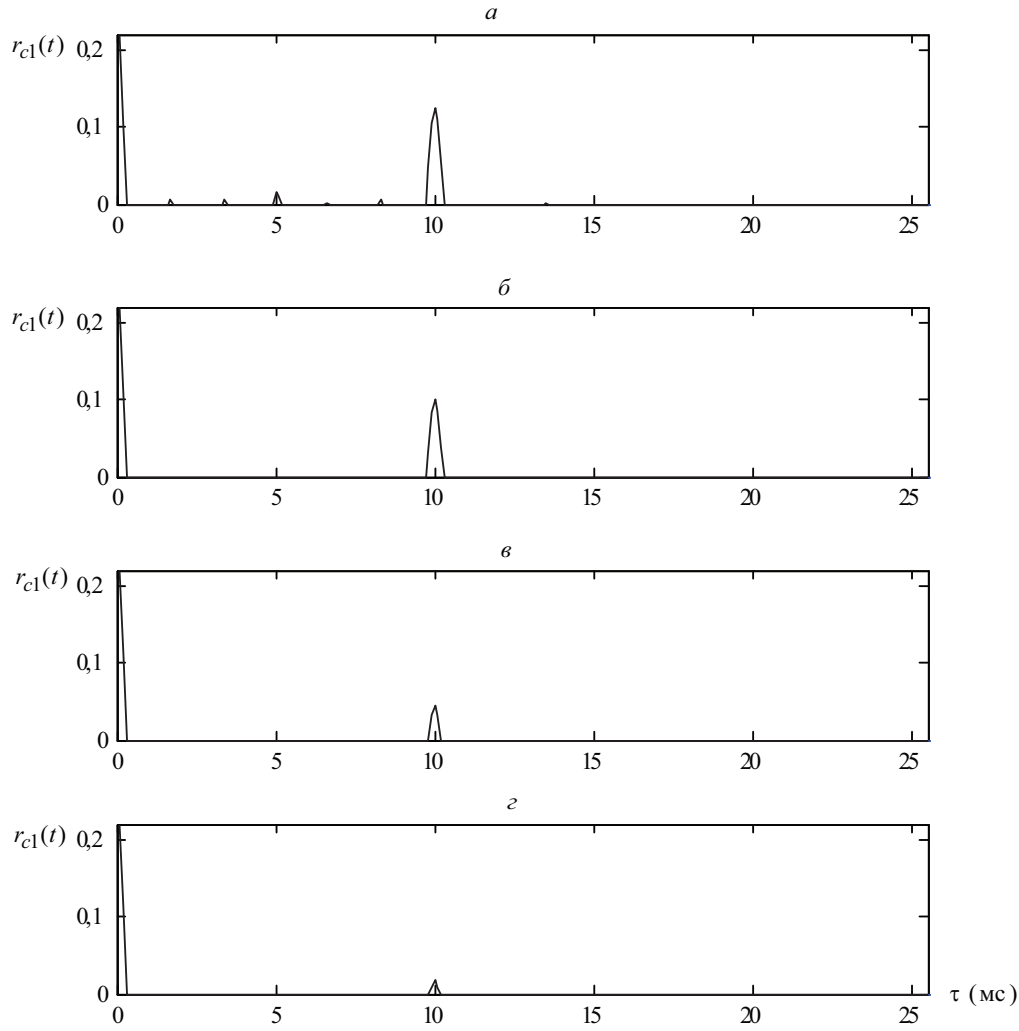

Рис. 6. Демонстрация устойчивости предлагаемого способа к изменениям амплитуды сигнала для случаев: *а* — $0 = 1$, *б* — $a = 0{,}5$, *в* — $a = 0{,}25$, *г* — $a = 0{,}125$.

ства измерения основного тона с помощью автокорреляционного [10] и предложенного способов проводился на примере 240 образцов шести естественных гласных: «а», «у», «э», «о», «ы», «и». Образцы были собраны от пяти мужчин и пяти женщин, четырехкратно произносивших короткие речевые фразы. Запись образцов производилась через динамический микрофон МД-71 с помощью 16-разрядной звуковой карты при частоте квантования 22,05 кГц. Образцы гласных имели длительность $\Delta T = 23{,}2$ мс и включали по 512 дискретных отсчетов. При этом число ошибок измерения основного тона составили соответственно 28 и 8 для автокорреляционного и предложенного способов.

Для демонстрации нечувствительности метода к изменениям амплитуды речевого сигнала был использован фрагмент двухформантного синтетическо-

го гласного длительностью $\Delta T = 18$ мс при частоте дискретизации 10 кГц. Синтезированный гласный имел частоту основного тона $f_0 = 100$ Гц и частоты формант $F_1 = 600$ Гц и $F_2 = 830$ Гц. Параметр α, определяющий уровень клиппирования, был выбран равным 0,8.

Сглаживание $r_{0e}(\tau)$ выполнялось с помощью фильтра низких частот с симметричной импульсной характеристикой $h(n) = 0{,}25u_0(n-1) + 0{,}5u_0(n) + 0{,}25u_0(n+1)$, где $n = \ldots -2, -1, 0, 1, 2, \ldots$, $u_0(n) = 1$ при $n = 0$ и $u_0(n) = 0$ при $n \neq 0$. Поэтому вычисление свертки сводилось к суммированию трех взвешенных спектральных отсчетов. На рис. 5 предложенный способ поясняется на примере гласного с постоянной амплитудой на протяжении сегмента длительностью $\Delta T = 18$ мс.

На рис. 6 продемонстрирована устойчивость предложенного способа при линейном убывании амплитуды гласного для случаев разной скорости убывания амплитуды сигнала, что обеспечивалось выбором $a = 1$, 0,5, 0,25, 0,125.

Из приведенных рисунков можно видеть, что предложенный способ обработки автокорреляционной функции позволяет подчеркнуть ее пик на $\tau = 1/f_0$, связанный с периодом сигнала T_0, как в случае речевого сигнала с постоянной амплитудой, так и при изменениях его амплитуды на интервале анализа ΔT. При этом во всех случаях пик у $r_{c1}(\tau)$ на $\tau = 1/f_0$ выражен более четко в сравнении с другими пиками, нежели у автокорреляционной функции $r_0(\tau)$.

4. Заключение

Таким образом, приведенные выше результаты исследования позволяют заключить, что предложенный способ обработки функции автокорреляции позволяет подчеркнуть ее пик на периоде сигнала T_0 и уменьшить число ошибок измерения основного тона речевого сигнала при наличии амплитудных вариаций сигнала на интервале анализа ΔT.

СПИСОК ЛИТЕРАТУРЫ

1. *Фант Г.* Акустическая теория речеобразования. М.: Наука, 1964.
2. *Hess W.* Pitch determination of signals. Berlin: Springer – Verlag, 1983.
3. *Маркел Д.Д., Грэй А.Х.* Линейное предсказание речи. М.: Связь, 1980.
4. *Чайлдерс Д.Дж., Скиннер Д.П., Кемерейт Р.Ч.* Кепстр и его применение при обработке данных // ТИИЭР. 1977. Т. 5. № 10. С. 5–23.
5. *Колоколов А.С.* Измерение основного тона речевого сигнала // АиТ. 2003. № 8. С. 122–134.
   Kolokolov A.S. Measuring the Fundamental Tone of Voice Signal // Autom. Remote Control. 2003. V. 64. No. 8. P. 1310–1320.
6. *Stephan D.E., Carolin T.I., Volker H.* Robust fundamental frequency estimation in an auditory model // AIA-DAGA. 2013. Merano. P. 271–274.
7. *Имамвердиев Я.Н., Сухостат Л.В.* Метод оценки периода основного тона с применением эмпирического вейвлет преобразования // Радіоелектроніка, інформатика, управління. 2015. № 2. С. 47–53.
8. *Aasha D.E., Ramesh Shweta, Kathuria Chhavi, Biswas Debdatta.* Comparative study of pitch estimation using harmonic product spectrum derived from DFT, DCT, Haar and KL transforms // Int. Pure Appl. Math. 2017. V. 115. No. 6. P. 403–408.

9. *Баронин С.П.* Автокорреляционный метод выделения основного тона речи / Сб. тр. Гос. НИИ Мин. связи СССР. Вып. 3(24). М., 1961. С. 93–102.
10. *Рабинер Л.Р., Шафер Р.В.* Цифровая обработка речевых сигналов. М.: Радио и связь, 1981.
11. *Sondhi M.M.* New methods of pitch extraction // IEEE Trans. Audio Electroacoust. 1968. V. AU-16. No. 2. P. 262–266.
12. *Колоколов А.С., Любинский И.А., Мещеряков А.Ю.* Измерение основного тона речевого сигнала на основе его автокорреляционной функции // Наукоемкие технологии. 2012. Т. 13. № 5. С. 26–29.
13. *Колоколов А.С., Павлова М.И.* Способ обработки функции автокорреляции для измерения основного тона речевого сигнала // Патент на изобретение № 2559710. Решение о выдаче от 27.05.2015.

Статья представлена к публикации членом редколлегии О.Н. Граничиным.

Поступила в редакцию 03.04.2018

После доработки 29.06.2018

Принята к публикации 08.11.2018