Вычисление мгновенных гармонических параметров речевого сигнала

И.С. Азаров,
аспирант,

А.А. Петровский,
доктор технических наук, профессор

В данной работе предлагается способ определения мгновенных параметров (частоты, амплитуда и фазы) основного тона и его гармоник в вокализованном речевом сигнале. Рассматривается алгоритм, для определения данных параметров, основанный на применении частотно-модулированных узкополосных фильтров, позволяющих представить выходной сигнал в виде однокомпонентной периодической функции. Показан способ синтеза данных фильтров. Полученные результаты могут быть использованы при решении задач синтеза и кодирования речевых сигналов, распознавания и конверсии речи, а также для редактирования акустических шумов в речевых сигналах.

1. Введение

Синусоидальная модель представления речевого сигнала известна с начала 80-х годов [1] и с тех пор успешно применяется в задачах кодирования и анализа речи [2-3]. В её основе лежит метод представления сигнала в виде суммы периодических тригонометрических функций (синусоид) различной амплитуды, частоты и фазы. Одним из инструментов, часто используемых для представления сигнала в такой форме, является преобразование Фурье или какая-либо его модификация. Как известно, в дискретном преобразовании Фурье количество синусоид, необходимых для представления исходного сигнала, равно числу точек преобразования. Но как показали исследования, достаточно использовать всего несколько, определённым способом выбранных синусоид, для того, чтобы описать основную (имеющую большую долю энергии) часть речевого сигнала. Эти «значащие» синусоиды располагаются в частотной области гармонически, т.е. через равные интервалы, причём ширина интервалов равна частоте основного тона. Модель, использующая для представления сигнала синусоиды кратной частоты, называется гармонической. Оставшаяся часть сигнала, которая не может быть описана при помощи данной модели, называется шумовой [4,5].

Представление сигнала в форме гармоники+шум эффективно используется во многих речевых приложениях [6,7]. В качестве примеров можно привести задачи синтеза, кодирования, распознавания речи, идентификации личности диктора, конверсии и восстановления повреждённых записей речевых сигналов.

Точная сепарация речевого сигнала на периодическую и шумовую составляющие является сложной задачей, при решении которой принимают ряд допущений. Так, большинство анализаторов [2,3] рассматривают речевой сигнал как стационарный внутри некоторого временного интервала (сегмента), что позволяет применять упрощённые методы и уменьшить сложность вычислений. Однако, принимая амплитуду и частоту гармонических компонент сигнала постоянными, невозможно разделить сигнал достаточно точно, что проявляется, например, в кодерах в виде различных звуковых артефактов.

Существуют подходы анализа, допускающие линейную частотную модуляцию частоты основного тона внутри каждого сегмента сигнала, которая определяется, например, при помощи гармонического преобразования [8]. Однако принимается, что частоты гармоник строго кратны частоте основного тона, а их амплитуды постоянные на протяжении каждого из сегментов сигнала.

Целью данной работы является нахождение простого и эффективного способа для определения мгновенных значений амплитуды, частоты и фазы гармоник основного тона, допуская при этом, что эти параметры изменяются нелинейно во времени с каждым новым отсчётом речевого сигнала.

Как показано в [9], можно достаточно просто вычислить мгновенную частоту и амплитуду, имея в своём распоряжении четыре последовательно расположенных отсчёта сигнала. Там же были предложены соответствующие алгоритмы (известные как DESA и DESA-2), позволяющие с высокой точностью определить значения этих параметров. Однако существует ряд ограничений, одно из которых — неприменимость алгоритмов для многокомпонентных сигналов без предварительной узкополосной фильтрации.

Предлагается способ синтеза узкополосного частотно-модулированного фильтра, а также алгоритм определения амплитуд, частот и фаз основного тона и его гармоник, учитывающий нестационарность и гармоническую структуру речевого сигнала.

2. Гармоническая модель речевого сигнала. Постановка задачи

Как сказано выше, речевой сигнал может быть представлен в виде суммы двух основных составляющих: периодической и шумовой. Таким образом, дискретный речевой сигнал можно записать в виде:

$$s(n) = \sum_{k=1}^{K} A_k(n)\cos\varphi_k(n) + r(n) \ , \qquad (1)$$

где A_k — мгновенная амплитуда k-ой гармоники, K — число гармоник, присутствующих в сигнале, $r(n)$ — шумовая компонента, $\varphi_k(n)$ — мгновенная фаза k-ой гармоники, которую можно вычислить, зная мгновенную частоту и начальную фазу:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0), \qquad (2)$$

где f_k — мгновенная частота k-ой гармоники, F_s — частота дискретизации и $\varphi_k(0)$ — начальная фаза k-ой гармоники. В гармонической модели частоты гармоник в каждый момент времени принимаются кратными частоте основного тона, т.е. выполняется следующее соотношение: $f_k = k f_0$, где k — номер гармоники, а f_0 — частота основного тона.

В данной работе считается, что расположение гармоник основного тона в частотной области удовлетворяют соотношению, близкому к гармоническому

$$\left| f_k - k f_0 \right| < f_{tr} \qquad (3)$$

где f_{tr} — максимальная величина возможного отклонения частоты гармоники.

Оценкой точности определения мгновенных параметров (амплитуды частоты и фазы) может служить отношение «гармоники/шум» [10]: $HNR = 10 \lg E_h / E_r$, где E_h и E_r — энергии гармонической и шумовой компоненты соответственно. Гармоническая компонента может быть получена после определения её мгновенных параметров путём синтеза по формуле (1). Шумовая компонента речевого сигнала определяется путём вычитания из исходного сигнала синтезированной гармонической компоненты.

Как показано в [9], для определения мгновенных параметров сигнала необходимо произвести фильтрацию в узкой частотной области, содержащей искомый компонент. Для того, чтобы эффективно использовать данный фильтр при определении мгновенных гармонических параметров в речевом сигнале, необходимо выполнение следующих требований:

— обеспечивать фильтрацию в произвольной полосе частот, для чего импульсная характеристика фильтра должна быть аналитически выражена через его центральную частоту и ширину частотной полосы фильтра;

— иметь возможность аналитически представить выходной сигнал фильтра в виде периодической функции (синусоиды) с мгновенной частотой и амплитудой через отсчёты входного сигнала и полосу фильтрации, для непосредственного определения мгновенных гармонических параметров;

— импульсная характеристика фильтра должна быть непрерывной.

Последнее условие обеспечивает возможность синтезировать частотно-модулированный фильтр — фильтр, осуществляющий фильтрацию в области гармонических компонент.

3. Оценка мгновенных гармонических параметров речевого сигнала

3.1. Синтез узкополосного фильтра на основе дискретного преобразования Фурье

Для фиксированной частоты f преобразование Фурье

$$S(f) = \frac{1}{N} \sum_{n=0}^{N-1} s(n) e^{\frac{-j2\pi nf}{N}},$$

$$MAG[S(f)] = \sqrt{\mathrm{Re}\,S(f)^2 + \mathrm{Im}\,S(f)^2},$$

$$\phi[S(f)] = -\arctan\frac{\mathrm{Im}\,S(f)}{\mathrm{Re}\,S(f)}$$

можно рассматривать как фильтр, преобразовывающий исходный сигнал в синусоиду с постоянной амплитудой *MAG*, с частотой f и с определённой начальной фазой φ на сегменте сигнала из *N* отсчётов. Непрерывная импульсная характеристика данного фильтра запишется в виде:

$$h(n) = \cos(\frac{2\pi}{F_s} nf),$$

где *h(n)*— импульсная характеристика фильтра.

По аналогии можно построить фильтр, преобразовывающий исходный сигнал в сумму синусоид с частотами из заданного диапазона, импульсная характеристика которого будет иметь вид

$$h(n) = \int_{F_1}^{F_2} \cos(\frac{2\pi}{F_s} nf)df,$$

где F_1 и F_2 — начало и конец частотного диапазона соответственно. После преобразования и раскрытия неопределённости при $n = 0$ получим следующее выражение:

$$h(n) = \begin{cases} F_2 - F_1, & n = 0 \\ \frac{F_s}{n\pi}\cos(\frac{n\pi}{F_s}(F_2 + F_1))\sin(\frac{n\pi}{F_s}(F_2 - F_1)), & n \neq 0 \end{cases}$$

Сегмент выходного сигнала $\bar{s}(n)$ из *N* отсчётов представляет собой результат свёртки *s(n)* и *h(n)*:

$$\bar{s}(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi}\cos(\frac{(n-i)\pi}{F_s}(F_2 + F_1))\sin(\frac{(n-i)\pi}{F_s}(F_2 - F_1)). \qquad (4)$$

После соответствующих преобразований выражение (4) принимает вид:

$$\bar{s}(n) = A(n)\cos(\frac{2\pi}{F_s} nF_c) + B(n)\sin(\frac{2\pi}{F_s} nF_c), \qquad (5)$$

где

$$A(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi}\sin(\frac{2\pi}{F_s} F_\Delta(n-i))\cos(\frac{2\pi}{F_s} F_c i),$$

$$B(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi}\sin(\frac{2\pi}{F_s} F_\Delta(n-i))\sin(\frac{2\pi}{F_s} F_c i),$$

$$F_c = \frac{F_2 + F_1}{2}, \qquad F_\Delta = \frac{F_2 - F_1}{2}.$$

Для того, чтобы упростить запись импульсной характеристики фильтра и формулы (4), используется прямоугольное окно. Очевидно, что для выполнения вычислений на практике обязательно следует применять окно более близкое к оптимальному (эксперименты в данной работе проводились с окном Кайзера (параметр β = 5,658)). Использование непрямоугольного окна лишь незначительно изменяет формулу (4) (для этого *s(i)* следует заменить на *s(i)* · *w(i)*, где *w(i)* — оконная функция).

Таким образом, выходной сигнал полосового фильтра можно аналитически представить в виде синусоиды с амплитудной и частотной модуляцией:

$$\bar{s}(n) = C(n)\cos(\frac{2\pi}{F_s}F_c n + \alpha(n)) \ ,$$

где $C(n) = \sqrt{A^2(n) + B^2(n)}$, $\alpha(n) = \arctan(-\frac{B(n)}{A(n)})$.

Отсюда можно вычислить мгновенную частоту *F*, амплитуду *MAG* и фазу φ :

$$F(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi}F_s + F_c \ ,$$

$$MAG(n) = C(n) \ , \quad \phi(n) = 2\pi F_c n + \alpha(n) \ .$$

Для определения мгновенных параметров гармоники достаточно знать примерную её частоту F_c и ширину полосы фильтрации. Поскольку предполагается выполнение соотношения (2), F_c может быть приблизительно вычислена следующим образом: $F_c = k f_0$.

На *рис. 1* приведена амплитудно-частотная характеристика синтезированных фильтров для основного тона 210 Гц и 17-и его гармоник, ширина частотных полос составляет 37,5 Гц. В качестве иллюстрации на *рис. 2* показан синтезированный тестовый частотно-модулированный гармонический сигнал (частоты и амплитуды основного тона и его гармоник), а на *рис. 3* приведён результат анализа данного сигнала, полученный при помощи фильтров, синтезированных как описано выше.

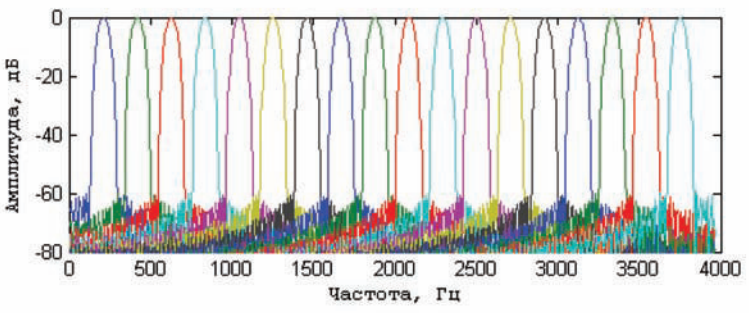


Рис.1. Пример амплитудно-частотной характеристики синтезированных фильтров

Анализ результатов моделирования *(рис. 3)* показывает, что если частота основного тона изменяется слишком быстро, то становится невозможным проследить частотные траектории гармоник, используя фильтр с постоянной полосой. Особенно сильно данный факт проявляется при работе с гармониками высоких порядков, поскольку величина модуляции кратна номеру гармоники.

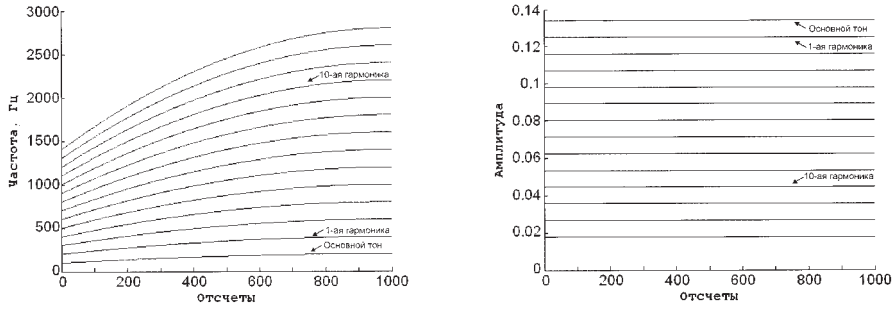


Рис.2. Частота (а) и амплитуда (б) основного тона и гармоник тестового сигнала

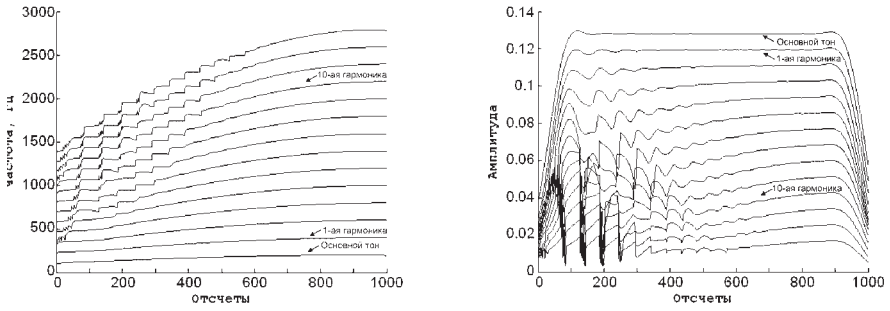


Рис. 3. Частота (а) и амплитуда (б) основного тона и гармоник, полученные из синтезированного сигнала при помощи анализа

3.2. Синтез узкополосного частотно-модулированного фильтра

Вокализованный речевой сигнал имеет частотные модуляции основного тона, что, как видно из рис. 3, усложняет процесс анализа. Традиционно для преодоления данной проблемы используется временное масштабирование [11]. Данная операция подразумевает получение из исходного сигнала *s(n)*, отсчёты которого расположены через равные интервалы времени, некоторого сигнала *s'(φ)*, отсчёты которого расположены через равные интервалы фазы основного тона. Таким образом, сигнал лишается частотных модуляций и может быть эффективно обработан тем или иным алгоритмом анализа. Этот способ, однако, имеет ряд недостатков: 1) качество сигнала может быть ухудшено при масштабировании, поскольку определить значения дискретного сигнала в произвольные моменты времени с достаточной точностью является сложной задачей; 2) для масштабирования сигнала используются дополнительные вычислительные ресурсы; 3) полученный результат требует дополнительной обработки для перехода обратно в частотно-временную область.

Синтезированные в данной работе фильтры имеют непрерывную импульсную характеристику, что позволяет выполнить временное масштабирование, интегрируя его непосредственно в фильтр, не масштабируя исходный сигнал. С учётом модуляции частоты основного тона выражение (5) для выходного сигнала фильтра примет вид:

$$\bar{s}(n) = A(n)\cos(\frac{2\pi}{F_s}\varphi^k(n)) + B(n)\sin(\frac{2\pi}{F_s}\varphi^k(n)) \ ,$$

где $\varphi^k(n) = (\sum_{i=0}^{n} F_0(n) - \sum_{i=0}^{N/2} F_0(n))k \ ,$

$F_0(n)$ — мгновенная частота основного тона, k — номер гармоники основного тона,

$$A(n)=\sum_{i=0}^{N-1}\frac{s(i)F_s}{(n-i)\pi}\sin\left(\frac{\pi}{F_s}F_\Delta(n-i)\right)\cos\left(\frac{2\pi}{F_s}\varphi^k(n)\right),$$

$$B(n)=\sum_{i=0}^{N-1}\frac{s(i)F_s}{(n-i)\pi}\sin\left(\frac{\pi}{F_s}F_\Delta(n-i)\right)\sin\left(\frac{2\pi}{F_s}\varphi^k(n)\right),$$

F_Δ — ширина полосы фильтрации.

Мгновенная частота, амплитуда и фаза определяются соответственно:

$$F(n)=\frac{\alpha(n+1)-\alpha(n)}{2\pi}F_s+F_0\cdot k, \qquad (6)$$

$$MAG(n)=C(n), \qquad \phi(n)=2\pi F_0 kn+\alpha(n).$$

Таким образом, полоса фильтра частотно модулируется в зависимости от частоты основного тона, что позволяет сохранить точность вычислений в случае его быстрого изменения. На *рис. 4*. приведён пример импульсной характеристики частотно-модулированного фильтра длиной в 321 отсчёт (ширина полосы фильтрации 75 Гц, центр полосы от начала до конца импульсной характеристики изменяется от 320 Гц до 450 Гц).

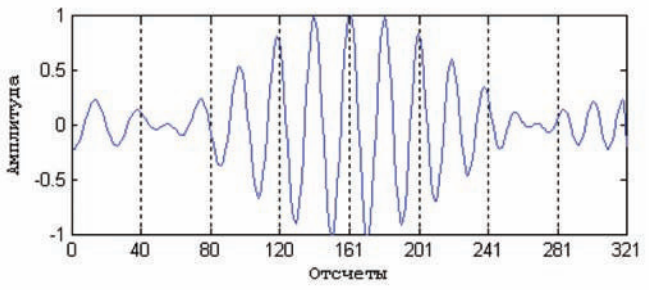


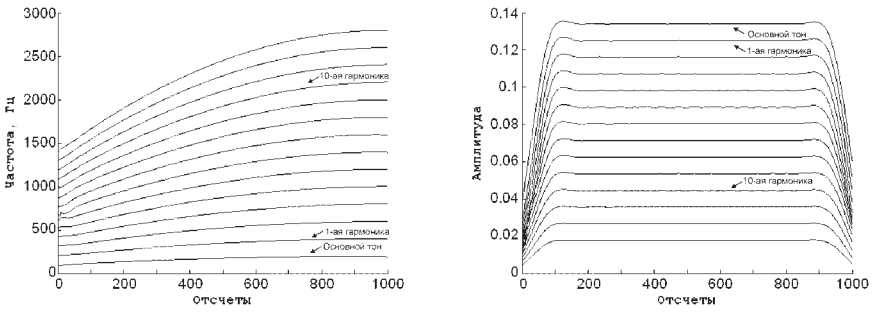
Рис.4. Импульсная характеристика частотно-модулированного фильтра

Результаты моделирования по оценке частотных траекторий гармоник с применением частотно-модулированного фильтра приведены на *рис. 5*. Сравнительный анализ со способом, основанным на фильтрации с постоянной частотной полосой *(рис. 3)*, показывает явное преимущество метода на базе частотно-модулированного фильтра по точности оценки частотных траекторий гармоник независимо от скорости изменения частоты основного тона.

4. Декомпозиция речевого сигнала на гармоническую и шумовую составляющие

4.1. Алгоритм определения гармонических параметров речевого сигнала

Для точной сепарации речевого сигнала на гармоническую и шумовую составляющие необходимо знание контура частоты основного тона. Это позволяет оценивать число фильтров и их расположение в частотной области в каждый момент времени, а так же синтезировать их импульсные характеристики с учётом модуляции частоты основного тона. Контур частоты основного тона может быть получен при помощи узкополосного фильтра, описанного в данной работе. Перемещая центр частотной полосы фильтра на каждом отсчёте соответственно с полученной частотой на предыдущем, контур можно точно проследить на протя-



а) б)

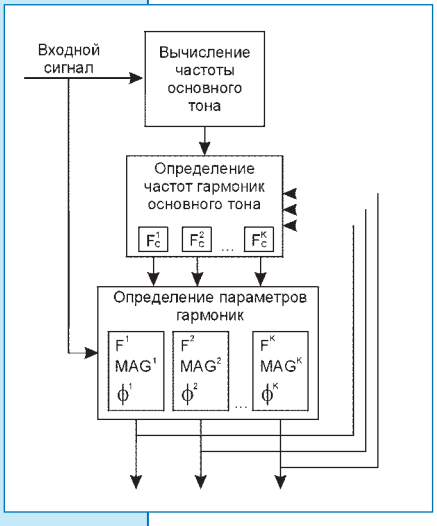
Рис.5. Частота (а) и амплитуда (б) основного тона и гармоник, полученные из синтезированного тестового сигнала (рис. 2) при помощи анализа с использованием частотно-модулированного фильтра

жении всего вокализованного отрезка сигнала, вокализованность которого определяется оценкой амплитуды основного тона. Если она превышает некоторое пороговое значение на рассматриваемом отсчёте, значит, данный отсчёт принадлежит вокализованной части сигнала. Тем не менее, требуется каким-либо образом получить частоту основного тона в начале вокализованного участка. Для этого используется методика, предложенная в [10], основанная на приблизительной оценке частоты при помощи автокорреляции с последующим уточнением оценки методом анализа-через-синтез.

Определяемые параметры — это тройка (амплитуда, частота, фаза) гармоник основного тона в каждый момент времени. Алгоритм состоит из следующих основных этапов:

1) оценка частоты основного тона;
2) синтез банка частотно-модулированных фильтров для данного момента времени;
3) вычисление мгновенных гармонических параметров из соотношения (5). Переход к пункту 2) либо выход (если достигнут последний отсчет на сегменте сигнала).

В пункте 2) алгоритма первоначально центры частотных полос фильтров принимаются равными произведению частоты основного тона на номер гармоники $F_c^k = F_0 k$, однако на последующих отсчётах полагается равной частоте гармоники с номером, полученной на предыдущем шаге, что обеспечивает последовательное приближение алгоритма.



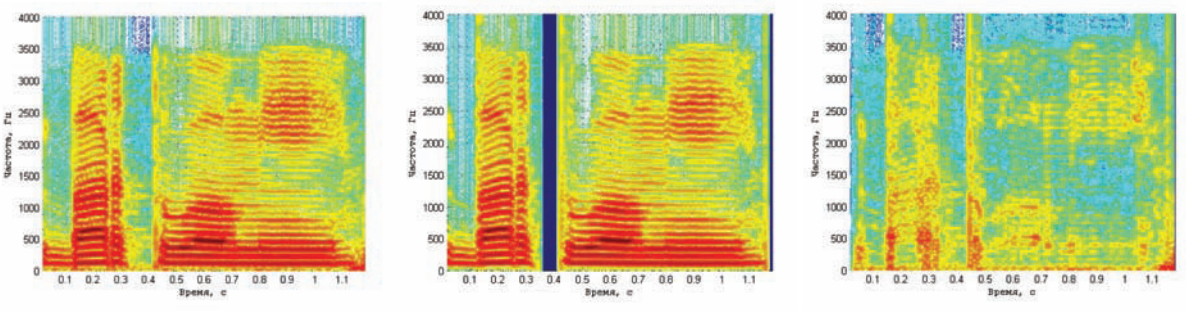
Общая схема алгоритма вычисления параметров представлена на *рис. 6*.

4.2. Экспериментальные результаты

В качестве эксперимента была выполнена сепарация речевого сигнала (запись мужского голоса), с частотой дискретизации 8кГц. Спектрограмма исходного сигнала, гармонической части и шумовой представлена на *рис.7*.

Исходный речевой сигнал содержит 28 гармоник основного тона, каждая из которых имеет амплитудную и час-

Рис. 6. Общая схема алгоритма определения гармонических параметров



а) *б)* *в)*

Рис.7. Спектрограмма исходного речевого сигнала (а), спектрограмма полученной гармонической части сигнала (б), спектрограмма полученной шумовой составляющей сигнала (в)

тотную модуляции. Как видно из *рис. 7 (б)*, синтезированная гармоническая часть сигнала содержит все 28 гармоник основного тона, сохраняя частотные и амплитудные контуры каждой из них. Полученная шумовая составляющая сигнала *(рис.6. (в)*) не содержит регулярных гармонических компонент и значительно ниже по уровню энергии, чем исходный речевой сигнал и его гармоническая компонента (отношение «гармоники/шум» (*HNR* = 21,33 дБ). На *рис. 8.* показаны временные формы полученных сигналов.

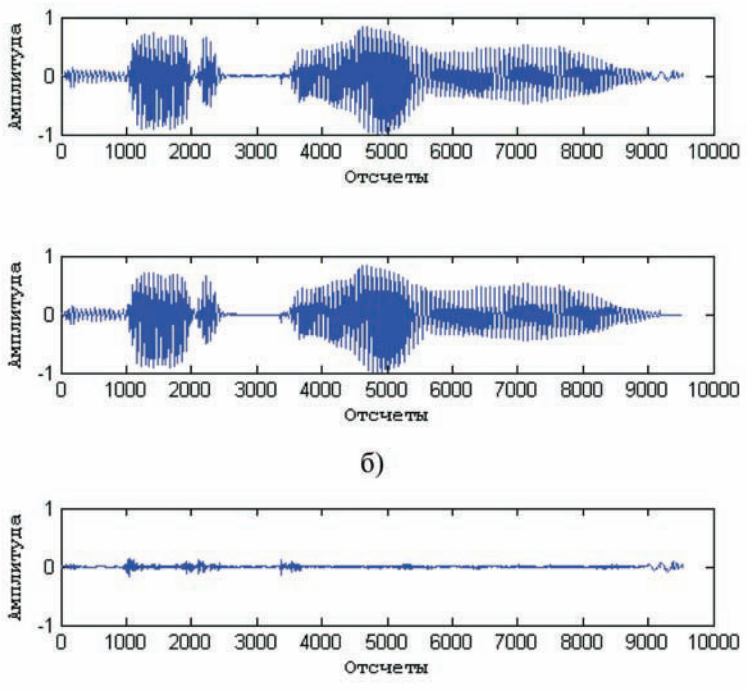


Рис. 8. Результат сепарации сигнала: исходный сигнал (а), периодическая часть (б), шумовая часть (в) (HNR = 21,33 дБ)

5. Выводы

Предложен способ определения мгновенных параметров (амплитуды, частоты, фазы) основного тона речевого сигнала и его гармоник при помощи цифрового фильтра с аналитической, непрерывной импульсной характеристикой. Тройки параметров определяются в результате фильтрации речевого сигнала узкополосным частотно-модулированным фильтром. Границы частотных областей фильтров вычисляются последовательно путём отслеживания частотных траекторий гармонических компонентов с учётом модуляции частоты основного тона.

Данный метод обеспечивает высокую точность сепарации гармонического сигнала, является легко реализуемым и может быть эффективно использован в приложениях, требующих разделения речевого сигнала на периодическую и шумовую компоненты.

Литература

1. *McAulay R.J., Quatieri T.F.* «Speech analysis/synthesis based on a sinusoidal representation» IEEE Trans. On Acoustics, Speech and Signal Process., 1986, vol. 34, no. 4, pp.744-754.
2. *Spanias A.S.* „Speech coding: a tutorial review», Proc. of the IEEE, 1994, vol. 82, no. 10, pp. 1541-1582.
3. *McAulay R.J., Quatieri T.F.* „Sinusoidal Coding» in Speech Coding and Synthesis (W. Klein and K. Palival, eds.), Amsterdam: Elsevier Science Publishers, 1995, pp. 121-176.
4. *George E.B., Smith M.J.T.* „Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model», IEEE Trans. on Speech and Audio Process., 1997, vol. 5, no. 5, pp. 389-406.
5. *Yegnanarayana B., d'Alessandro C., Darsions V.,* «An Iterative Algorithm for Decomposition of Speech Signals into Voiced and Noise Components», IEEE Trans. on Speech and Audio Coding, 1998, vol. 6, no. 1, pp. 1-11.
6. *Stylianou Y.* «Applying the harmonic plus noise model in concatenative speech synthesis» IEEE Trans. Speech, Audio Process., 2001, vol. 9, no. 1, pp.21-29.
7. *Zavarehei E., Vaseghi S., Yan Q.* «Noisy speech enhancement using harmonic-noise model and codebook-based post-processing», 2007, vol. 15, no. 4, pp.1194-1203.
8. *Zubrycki P., Petrovsky A.* Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform // 15th European Signal Process. Conf., (EUSIPCO-2007), Poznan, 2007, pp.2336-2340.
9. *Maragos P., Kaiser J. F., Quatieri T. F.*, «Energy Separation in Signal Modulations with Application to Speech Analysis», IEEE Trans. on Signal Process., 1993, vol. 41, no. 10, pp. 3024-3051.
10. *Pavlovets A., Kien T., Zubrycki P., Petrovsky A.* Speech analysis-synthesis based on the PTDFT for voice conversion. in Proc. of the 2007 Intern. TICSP Workshop on Spectral Methods and Miltirate Signal Processing, (SMMSP'2007), Moscow, 2007. — Tampere International Center for Signal Processing, TICSP Series #37, Tampere, 2007. — pp.203-210.
11. *Resch B., Nilsson M., Ekman A., Klejin W.B.* «Estimation of the Instantaneous Pitch of Speech», IEEE Trans. on Audio, Speech, and Lang. Process., 2007, vol. 15, no. 3, pp. 813-822.

Азаров Илья Сергеевич,

Аспирант,
Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»
Белорусский государственный университет, механико-математический факультет, отделение математической электроники.
Математик, математик-системотехник.
Область интересов: цифровая обработка сигналов, анализ/сжатие речевых сигналов.

Петровский Александр Александрович,

Доктор технических наук, профессор
Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники» (бывший Минский радиотехнический институт), кафедра Электронных вычислительных средств
Специальность — «Электронные вычислительные машины».
Главные научные интересы лежат в области цифровой обработки сигналов речи и звука для целей компрессии, распознавания, редактирования шума, а также проектирование проблемно-ориентированных средств вычислительной техники реального времени для систем мультимедиа.
Профессор Петровский А.А. является членом НТО РЭС им. А.С.Попова, IEEE, EURASIP, AES.