И. С. Азаров, М. И. Вашкевич, Д. С. Лихачев, А. А. Петровский, Изменение частоты основного тона речевого сигнала на основе гармонической модели с нестационарными параметрами, *Тр. СПИИРАН*, 2014, выпуск 32, 5–26

УДК 004.934

И.С. Азаров, М.И. Вашкевич, Д.С. Лихачев, А.А. Петровский

ИЗМЕНЕНИЕ ЧАСТОТЫ ОСНОВНОГО ТОНА РЕЧЕВОГО СИГНАЛА НА ОСНОВЕ ГАРМОНИЧЕСКОЙ МОДЕЛИ С НЕСТАЦИОНАРНЫМИ ПАРАМЕТРАМИ

Азаров И.С., Вашкевич М.С., Лихачев Д.С., Петровский А.А. **Изменение частоты основного тона речевого сигнала на основе гармонической модели с нестационарными параметрами.**

Аннотация. В статье предлагается решение задачи изменения частоты основного тона речевого сигнала. Необходимость решения данной задачи возникает во многих речевых приложениях таких как конверсия голоса, коррекция акцента, обеспечение конфиденциальности диктора и др. Разработанная схема обработки вокализованной части речевого сигнала основывается на гармонической модели с нестационарными (изменяющимися в каждый момент времени) параметрами. Для повышения частотного разрешения модели оценка параметров выполнятся при помощи узкополосной фильтрации в искривленном масштабе времени, согласованном с контуром мгновенной частоты основного тона. На основании субъективной оценки результатов показано, что разработанный способ обеспечивает высокую натуральность и разборчивость синтезированной речи и может применяться как в широкополосных так и в узкополосных каналах связи с различными стандартами кодирования (в том числе с кодеками G.711 и GSM).

Ключевые слова: гибридная модель речевого сигнала, оценка основного тона, изменение просодических характеристик речи.

Azarov E., Vashkevich M., Likhachov D., Petrovsky A. **Pitch modification of speech signal using harmonic model with time-varying parameters.**

Abstract. The paper presents a solution to the problem of pitch modification of speech. The problem occurs in different speech processing applications such as voice conversion, accent correction, hiding speaker's personality and other. Developed processing scheme for voiced part of speech is based on the harmonic model with nonstationary (time-varying) parameters. In order to improve frequency resolution of the model parameters are extracted using narrowband filtering in warped time domain aligned to instantaneous pitch frequency. Using subjective listening tests it is shown that developed system provides high naturalness and intelligibility of reconstructed speech and can be applied to wideband and narrowband communication channels with various coding standards (including G.711 and GSM).

Keywords: hybrid speech modeling, pitch estimation, prosody modification.

1. Введение. Эффект изменения частоты основного тона звукового сигнала может достигаться различными методами. Самым простым из них является изменение скорости воспроизведения, что приводит к смещению частоты всех составляющих сигнала. Однако, во-первых, это приводит к изменению длительности сигнала и потому не может быть использовано в приложениях, работающих в реальном масштабе времени, а во-вторых, это сильно искажает тембр голоса. Одним из наиболее популярных альтернатив является фазовый вокодер [1] и разнообразные методы на его основе, выполняющие смеще-

ние компонент сигнала в частотной области при помощи прямого и обратного преобразования Фурье. Применение фазового вокодера позволяет изменять высоту звучания сигнала без изменения длительности и сохранять тембр путем коррекции спектральной огибающей. Данный подход может быть использован для любых звуковых сигналов. Вокодер может быть реализован в режиме реального времени, поскольку выполняет обработку входного сигнала последовательно фрейм за фреймом. При применении его к речи и певческому голосу метод имеет существенные ограничения: 1) при обработке не используется какая-либо модель голосообразования, что при значительном изменении частоты основного тона приводит к неестественному звучанию; 2) выделяемые частотные составляющие сигнала не соответствуют гармоникам основного тона, что приводит к потере натуральности и звонкости голоса; 3) сигнал не разделяется на вокализованный и невокализованный, что приводит к неестественному звучанию некоторых звуков (в частности невокализованных шипящих ‘с’, ‘ш’, а так же смешанных, частично вокализованных, ‘з’ и ‘ж’).

Для повышения натуральности звучания и расширения доступного диапазона изменения основного тона необходимо использовать более сложные решения, основанные на гибридной (детерминистской/стохастической) модели речевого сигнала. Гибридная обработка звуковых сигналов предложена в работах [2,3], где используется три раздельные составляющие: периодическая, шумовая и транзиентная. Поскольку при изменении основного тона голоса обрабатывается только вокализованная (детерминистская) часть сигнала, в данном случае нет необходимости различать между собой шумовые и транзиентные составляющие – вместе их можно отнести к невокализованной (стохастической) части сигнала [4,5,6].

В настоящей работе предлагается решение задачи изменения основного тона, основанное на гибридном представлении речевого сигнала. Частотно-временной анализ выполняется с учетом контура частоты основного тона, что позволяет выделять параметры отдельных гармоник и разделять сигнал на вокализованную и невокализованную составляющие в частотной области. Моделирование вокализованного сигнала выполняется при помощи нестационарных (т.е. изменяющихся во времени) параметров. В экспериментальной части работы приводятся результаты субъективной оценки качества обработанной речи. С целью оценить применимость метода в существующих каналах связи для экспериментов использовались как широкополосные, так и узкополосные речевые сигналы с применением кодеков G.711 и GSM.

2. Гибридная модель и общая схема обработки речевого сигнала. Речевой сигнал представляется в виде суммы двух составляющих: вокализованной и невокализованной. Для описания вокализованной части сигнала используется синусоидальная модель [7,8]:

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

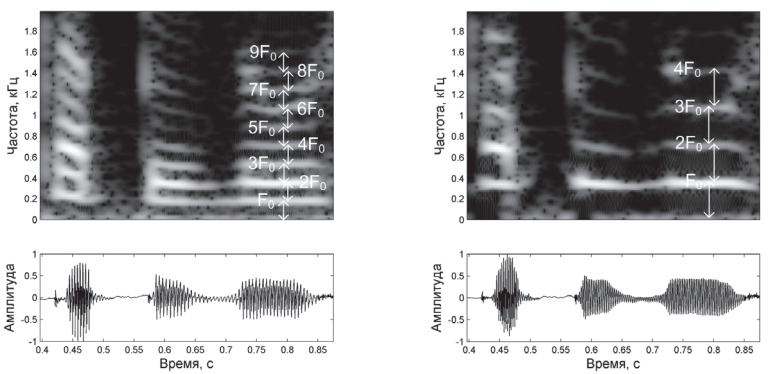
где $A_k(n)$ – мгновенная амплитуда -ой гармоники, K – число гармоник, $\varphi_k(n)$ – мгновенное значение фазы k-ой гармоники, $r(n)$ – шумовая составляющая сигнала. Мгновенная частота $f_k(n)$ связана с мгновенной фазой следующим соотношением:

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

где F_s – частота дискретизации и $\varphi_k(0)$ – начальная фаза -ой гармоники. Приближенно можно считать, что частота каждой гармоники является кратной частоте основного тона (как показано на рисунке 1), т.е.

$$f_k(n) \approx F_0(n)k,$$

где $F_0(n)$ - основной тон.



а б

Рис. 1. Изменение основного тона речевого сигнала;
а – исходный речевой сигнал; б – обработанный сигнал

Частота основного тона определяет высоту звучания голоса. Контур частоты основного тона (его изменение в зависимости от времени на большом протяжении) определяет интонацию речи.

Для того чтобы сохранить исходный тембр диктора при изменении основного тона должна сохранятся спектральная огибающая, которая формируется из мгновенных амплитуд гармоник основного тона. Спектральная огибающую можно рассматривать как функцию от номера отсчета и частоты $E(n, f)$, которая принимает значения мгновенных амплитуд гармоник основного тона в соответствующих точках $E(n, f_k(n)) = A_k(n)$. Для произвольных n и f функция вычисляется путем линейной интерполяции ближайших к ним амплитудных значений.

Синтез голоса с модифицированным контуром частоты основного тона может быть выполнен по следующей формуле:

$$s(n) = \sum_{k=1}^{K} E(n, \bar{F}_0(n)k)\cos \bar{\varphi}_k(n) + r(n),$$

где фазы гармонических компонентов $\bar{\varphi}_k(n)$ рассчитываются в соответствии с новым контуром частоты основного тона $\bar{F}_0(n)$ следующим образом:

$$\bar{\varphi}_k(n) = \sum_{i=0}^{n} \frac{2\pi \bar{F}_0(i)}{F_s} + \bar{\varphi}_k^{\Delta}(n).$$

Дополнительный фазовый параметр $\bar{\varphi}_k^{\Delta}(n)$ используется для сохранения относительных фаз гармоник по отношению к фазе частоты основного тона. Данный параметр вычисляется как

$$\bar{\varphi}_k^{\Delta}(n) = \varphi_k(n) - k\varphi_0(n).$$

Основной тон присутствует только в вокализованных сегментах речи, т.е. только тогда когда задействованы голосовые связки диктора. Такие звуки как 'а', 'о', 'ж' являются вокализованными, в то время как звуки 'с','ш','щ' являются невокализованными. Как было показано выше в спектре речевого сигнала вокализованность проявляется в виде спектральных компонент кратной частоты. Для того чтобы сохранить исходное качество звучания невокализованных звуков алгоритм изменения тона должен автоматически выделять области вокализованности и выполнять обработку только в этих областях. Шумовая часть $r(n)$ выделяется из исходного сигнала вычитанием выделенных вокализованных звуков как показано на рисунке 2.

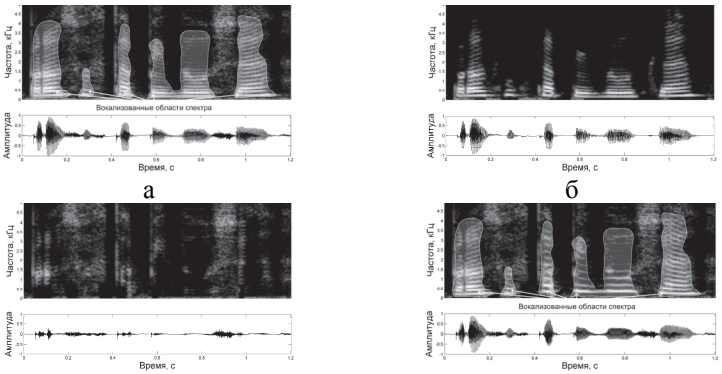


Рис. 2. Разделение сигнала на периодическую и стохастическую составляющие; а – исходный речевой сигнал; б – выделенная вокализованная компонента; в – сигнал-остаток $r(n)$; г – обработанный выходной сигнал

Алгоритм обработки сигнала можно кратко описать в виде последовательности действий, схематически изображенной на рисунке 3:

1) Определение частоты основного тона $F_0(n)$;

2) Преобразование речевого сигнала в параметрический вид т.е. оценка мгновенных гармонических параметров $A_k(n)$, $f_k(n)$ и $\varphi_k(n)$, $k = 1,2, \dots, K$;

3) Оценка вокализации каждой тройки гармонических параметров и отбор только тех, которые относятся к вокализованным областям спектра.

4) Синтез исходной вокализованной компоненты сигнала и ее вычитание из исходного речевого сигнала для получения шумовой составляющей $r(n)$.

5) Синтез вокализованной компоненты с измененным основным тоном и ее сложение с шумовой составляющей $r(n)$.

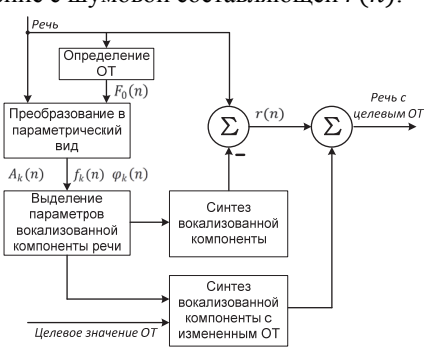


Рис. 3. Общая схема обработки сигнала

Оценка параметров модели выполняется с учетом контура мгновенной частоты основного тона. Входной речевой сигнал масштабируется во времени для того, чтобы обеспечить его стационарность, а затем выполняется узкополосная фильтрация, разделяющая гармоники основного тона. В результате фильтрации формируются аналитические (комплексные) сигналы, которые описываются при помощи параметров синусоидальной модели. На основе анализа смежных значений мгновенной частоты определяется степень вокализации каждой гармоники. Затем выполняется синтез квазипериодического сигнала по полученным параметрам, который вычитается из исходного речевого сигнала. В результате выполняется разделение речевого сигнала на вокализованную и невокализованную части. Одновременно синтезируется квазипериодический сигнал с целевой частотой основного тона, который складывается с полученной невокализованной частью.

Целевой контур основного тона формируется согласно требованиям конкретного приложения. Для экспериментов использовались заданные профили изменения исходного основного тона, приведенные в соответствующем разделе.

3. Оценка параметров речевого сигнала. Оценка основного тона выполняется при помощи алгоритма, изложенного в работах [9,10]. Особенностью алгоритма является возможность определения мгновенной частоты. Эта возможность достигается за счет использования специальной функции оценки периодичности аналогичной автокорреляционной функции [11], вычисляемой из мгновенных гармонических параметров субполосных составляющих сигнала:

$$\phi_{inst}(n,p) = \frac{\sum_{k=1}^{K} A_k^2(n)\cos(f_k(n)p)}{\sum_{k=1}^{K} A_k^2(n)}, \qquad (2)$$

где p – длина периода кандидата основного тона. В отличие от автокорреляционной функции, $\phi_{inst}(n,p)$ нечувствительна к любым изменениям частоты основного тона в окрестности отсчета n при условии, что используемые гармонические параметры получены достаточно точно.

Оценка параметров отдельных гармоник выполнятся при помощи узкополосной фильтрации. Для того, чтобы повысить частотное разрешение анализа масштаб времени сигнала изменяется согласованно с полученным контуром мгновенной частоты основного тона. Сигнал дискретизируется таким образом, чтобы на каждый период основного тона приходилось равное количество отсчетов N_{f0}.

Каждому входному отсчету речевого сигнала $s(n)$ ставится в соответствие фаза периода основного тона $\phi(n)$

$$\phi(n) = \sum_{i=0}^{n} f_0(i),$$

где $f_0(i)$ – нормализованная круговая частота основного тона в момент времени i. Новые моменты времени m в которые необходимо перерасчитать входной сигнал определяются как

$$m = \phi^{-1}\left(p/N_{f0}\right),$$

где p – это индекс отсчета в масштабированной временной области. Обратная функция $\phi^{-1}(\cdot)$ вычисляется при помощи линейной интерполяции ее известных значений для целых n как показано на рисунке 4.

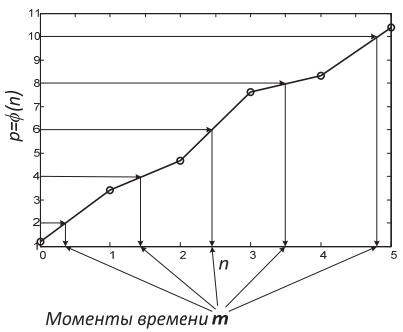


Рис. 4. Расчет новых моментов времени для масштабирования сигнала

Вычисление значений сигнала в заданный момент времени выполняется по теореме Котельникова (рисунок 5), согласно которой

$$s(t) = \sum_{n=-\infty}^{\infty} s(nT)\,\mathrm{sinc}\left(\frac{t-nT}{T}\right), \qquad (3)$$

где $\mathrm{sinc}(x) = \frac{\sin \pi x}{\pi x}$, T – интервал дискретизации, а $x(nT)$ – значения сигнала в дискретные моменты времени nT. Для простоты в дальнейшем примем $T = 1$ и вместо $s(nT)$ будем писать $s(n)$.

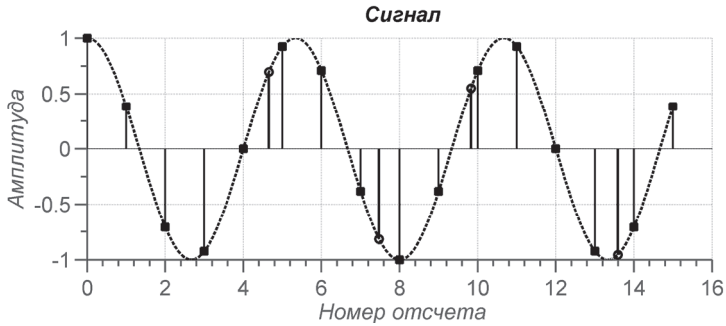


Рис. 5. Вычисление значений сигнала в произвольные моменты времени (пунктирная линия – непрерывный сигнал; квадратный маркер – сигнал дискретизированный с равным временным интервалом; круглый маркер – значения сигнала в моменты времени не кратные интервалу дискретизации)

Выражение (3) нельзя использовать на практике из-за суммирования в бесконечных пределах. Поэтому выбирается конечное число точек сигнала N_{pt} предшествующих моменту t и N_{pt} последующих точек:

$$s(t) = \sum_{n=-N_{pt}}^{N_{pt}} s(n)\,\mathrm{sinc}(t-n)\,w(t-n), \qquad (4)$$

где $w(\cdot)$ – оконная функция с центром симметрии в точке 0. Для каждого выходного отсчета временные отметки перерасчитываются таким образом, чтобы текущий момент t всегда попадал в диапазон от 0 до 1 как показано на рисунке 6. Это позволяет использовать таблицу с заранее рассчитанными значениями функции sinc. Использование табличных значений дает возможность существенно сократить вычислительные затраты.

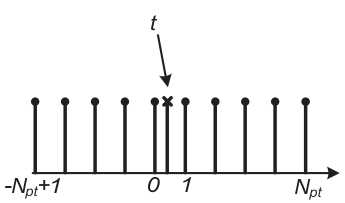


Рис. 6. Интерполяция сигнала функциями sinc со смещением от 0 до 1

После временного масштабирования основной тон становится постоянным и каждая гармоника частоты основного тона имеет фик-

сированное положение в частотном диапазоне. Для разделения гармоник и оценки их параметров используется ДПФ-модулированный банк фильтров анализа, который представляет собой совокупность фильтров $H_k(z)$, $0 \le k < M$, раскладывающих входной сигнал на M субполосных сигналов [12]. Амплитудно-частотные характеристики (АЧХ) фильтров формируются путем сдвига в частотной области АЧХ фильтра-прототипа. В результате полоса частот от нуля до частоты дискретизации перекрывается гребенкой из M фильтров, как показано на рисунке 7.

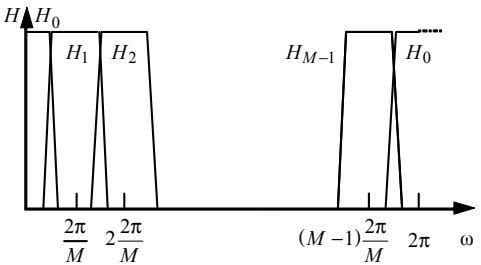


Рис. 7. АЧХ каналов ДПФ-модулированного банка фильтров

Число каналов в банке выбирается равным N_{f0}, поскольку в этом случае центральная частота каждого канала совпадает с положением соответствующей гармоники масштабированного сигнала $s(m)$. Импульсные характеристики фильтров анализа определяются выражением:

$$h_k(n) = h(n)W_M^{-kn}, \qquad k = 0,1,\dots M-1, \tag{5}$$

где $h(n)$ – фильтр-прототип нижних частот с нормированной частотой среза π/M, $W_M = e^{-j2\pi/M}$. Передаточная функция каждого фильтра анализа выражается через передаточную функцию фильтра-прототипа следующим образом:

$$H_k(z) = \sum_{n=0}^{N-1} h_k(n)z^{-n} = \sum_{n=0}^{N-1} h(n)(zW_M^{-k})^{-n} = H(zW_M^{-k}). \tag{6}$$

Фильтрация сигнала с использованием импульсных характеристик (5) является весьма неэффективной и приводит к большой вычислительной нагрузке. Более эффективный способ фильтрации можно получить, если воспользоваться полифазным представлением фильтра-

прототипа, в котором импульсная характеристика $h(n)$ перегруппирована в M подпоследовательностей $e_l(n)$

$$H(z) = \sum_{l=0}^{M-1} z^{-l} E_l(z^M), 0 \le l < M, \tag{7}$$

где

$$E_l(z) = \sum_{n=0}^{N/M-1} e_l(n) z^{-n} = \sum_{n=0}^{N/M-1} h(l + Mn) z^{-n}.$$

Далее объединим выражения (6) и (7). С учетом ва $W_M^{kM} = 1$ получим

$$H_k(z) = \sum_{n=0}^{M-1} z^{-l} W_M^{-lk} E(z^M W_M^{kM}) = \sum_{l=0}^{M-1} z^{-l} W_M^{-kl} E_l(z^M). \tag{8}$$

Уравнение (8) можно записать в виде произведения вектора строки на вектор-столбец

$$H_k(z) = \begin{bmatrix} 1 & W_M^{-k} & W_M^{-2k} & \dots & W_M^{-(M-1)k} \end{bmatrix} \begin{bmatrix} E_0(z^M) \\ z^{-1} E_1(z^M) \\ z^{-2} E_2(z^M) \\ \vdots \\ z^{-(M-1)} E_{M-1}(z^M) \end{bmatrix}.$$

Все M уравнений (для каждого H_k) в матричной форме можно записать следующим образом:

$$\begin{bmatrix} H_0(z) \\ H_1(z) \\ H_2(z) \\ \vdots \\ H_{M-1}(z) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W_M^{-1} & W_M^{-2} & \dots & W_M^{-(M-1)} \\ 1 & W_M^{-2} & W_M^{-4} & \dots & W_M^{-2(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_M^{-(M-1)} & W_M^{-2(M-1)} & \dots & W_M^{-(M-1)^2} \end{bmatrix} \begin{bmatrix} E_0(z^M) \\ z^{-1} E_1(z^M) \\ z^{-2} E_2(z^M) \\ \vdots \\ z^{-(M-1)} E_{M-1}(z^M) \end{bmatrix}. \tag{9}$$

Выражение (9) является основой для построения эффективной полифазной структуры ДПФ-модулированного банка фильтров, поскольку матрица $M \times M$ в правой части выражения является матрицей дискретного преобразования Фурье (ДПФ). Структура банка, использованная в настоящей работе, приведена на рисунке 8. При реализации ДПФ-модулированного банка фильтров порядок фильтра-прототипа выбирался равным: $N = kM + 1$, где k – некоторое целое число.

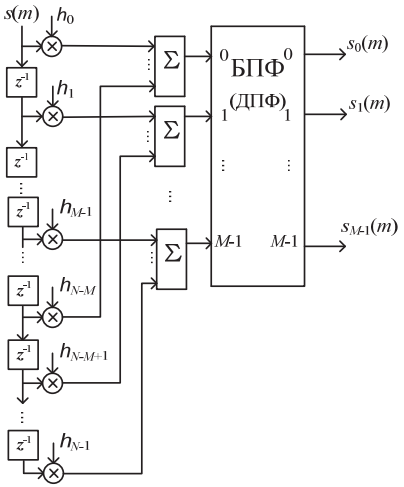


Рис. 8. Схема эффективной реализации ДПФ-модулированного банка фильтров анализа

Выходы банка фильтров $s_k(m)$ являются узкополосными аналитическими сигналами, которые интерпретируется как отдельные синусоидальные компоненты модели (1). Параметры синусоидальных компонент вычисляются с интервалом 5 мс при помощи следующих выражений:

$$A_k(m) = \sqrt{\mathrm{Re}^2\left(s_k(m+\Delta)\right) + \mathrm{Im}^2\left(s_k(m+\Delta)\right)}, \tag{10}$$

$$\varphi_k(m) = \arg s_k(m+\Delta), \tag{11}$$

$$f_k(m) = \frac{\arg s_k(m+\Delta+1) - \arg s_k(m+\Delta)}{2\pi}, \tag{12}$$

где $\Delta = \frac{N-1}{2}$ – групповая задержка банка фильтров.

4. Модификация гармонических параметров и синтез выходного сигнала. Полученные параметры гармонической ли $A_k(m)$, $\varphi_k(m)$ и $F_0(m)$ изменяются в соответствии с новым значением мгновенной частоты основного тона $\bar{F}_0(m)$. В каждый момент времени амплитуды гармоник входного сигнала A_k определены на сетке частот kF_0. Поскольку основной тон изменяется, но сохраняется

исходная спектральная огибающая, то необходимо перерасчитать амплитуды на новой частотной сетке $k\bar{F}_0$ как показано на рисунке 9. Перерасчет осуществляется при помощи линейной интерполяции.

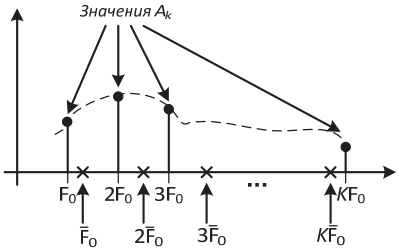


Рис. 9. Вычисление амплитудных значений гармоник новой частоты основного тона

Мгновенные фазовые значения изменяются таким образом, чтобы сохранить их непрерывность и синхронизацию фазы между гармониками:

$$\bar{\varphi}_k(m) = \bar{\varphi}_k(m-1) + \frac{\bar{F}_0(m)}{F_s} 2\pi\Delta m + \varphi_k(m) - kF_0(m),$$

где F_s – частота дискретизации, Δm – шаг (в отсчетах) между двумя последовательными оценками параметров модели.

Восстановление сигнала из параметрического представления выполняется путем синтезирования децимированных субполосных сигналов $\hat{s}_k(m)$ из измененных параметров для каждой гармоники. Затем эти сигналы подаются на вход банка фильтров синтеза, схема которого показана на рисунке 10. Через $\uparrow$S обозначен экспандер, добавляющий $S-1$ нулевых значений между последовательными отсчетами сигнала $\hat{s}_k(m)$. Предполагается, что число S является лем M. Через $F_k(z)$ на схеме обозначены фильтры синтеза, импульсные характеристики которых определяются следующим образом:

$$f_k(n) = h(n)W_M^{-kn},\ k = 0,1 \ldots M-1,\ n = 0,1 \ldots N-1. \qquad (13)$$

Прямая реализация данной схемы требует значительных вычислительных затрат. Её сложность можно уменьшить если учесть способ получения импульсных характеристик фильтров (13) и то, что большинство поступающих в фильтры отсчетов являются нулевыми. Полифазное представление фильтра-прототипа имеет вид:

$$H(z) = \sum_{l=0}^{M-1} z^{-(M-1-l)} R_l(z^M), \qquad 0 \le l \le M, \tag{14}$$

где $R_l(z) = E_{M-1-l}(z)$. Объединяя (13) и (14) получим:

$$F_k(z) = \sum_{l=0}^{M-1} z^{-(M-1-l)} W_M^{-(M-1-l)} R_l(z^M) = \sum_{l=0}^{M-1} z^{-(M-1-l)} W_M^{-kl} R_l(z^M).$$

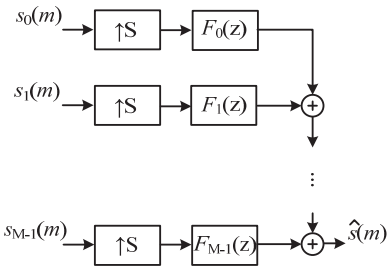


Рис. 10. Банк фильтров синтеза

Выражение для банка фильтров синтеза можно представить в матричной форме:

$$\hat{S}(z) = [F_0(z) \quad F_1(z) \quad F_2(z) \quad \dots \quad F_{M-1}(z)] \begin{bmatrix} S_0(z) \\ S_1(z) \\ S_2(z) \\ \vdots \\ S_{M-1}(z) \end{bmatrix}.$$

Его же можно записать в следующем виде:

$$\hat{S}(z) = [z^{-(M-1)} R_{M-1}(z^M) \quad z^{-(M-2)} R_{M-2}(z^M) \quad z^{-(M-3)} R_{M-3}(z^M) \quad \dots \quad R_0(z^M)] \times$$
$$\times \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & W_M^{-1} & W_M^{-2} & \dots & W_M^{-(M-1)} \\ 1 & W_M^{-2} & W_M^{-4} & \dots & W_M^{-2(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_M^{-(M-1)} & W_M^{-2(M-1)} & \dots & W_M^{-(M-1)^2} \end{bmatrix} \times \begin{bmatrix} S_0(z) \\ S_1(z) \\ S_2(z) \\ \vdots \\ S_{M-1}(z) \end{bmatrix}.$$

В этом выражении присутствует матрица размера $M \times M$ обратного дискретного преобразования Фурье (ОДПФ). В соответствии с этим строится эффективная полифазная структура ДПФ-модулированного банка фильтров синтеза, показанная на рисунке 11.

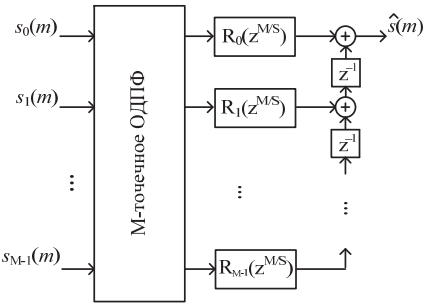


Рис. 11. Эффективная реализация банка фильтров синтеза

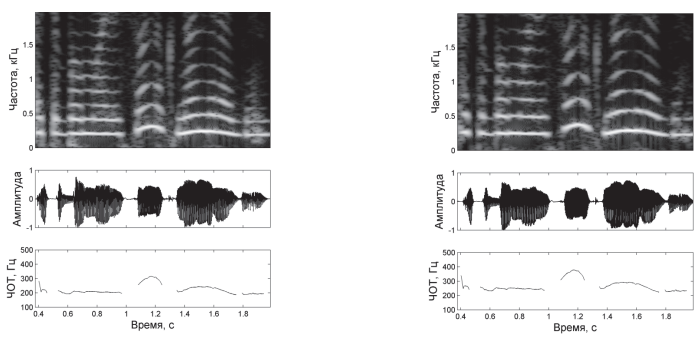
После реконструкции сигнала $\hat{s}(m)$ выполняется его обратное временное масштабирование с учетом целевого контура основного тона, в результате чего формируется выходной сигнал $\hat{s}(n)$ с постоянным шагом дискретизации.

5. Результаты экспериментов

5.1. Профили изменения контура основного тона. Для тестирования качества работы системы было разработано четыре эффекта (профиля) изменения частоты основного тона. Первый эффект заключался в повышении основного тона. Целевая частота основного тона $\bar{F}_0(n)$ и исходная $F_0(n)$ связываются соотношением:

$$\bar{F}_0(n) = F_0(n)k,\ k > 1,$$

где k – коэффициент на который происходит повышение тона. Высокая натуральность синтезированного голоса достигается, если коэффициент $k < 2$. На рисунке 12 показан пример повышения основного тона.



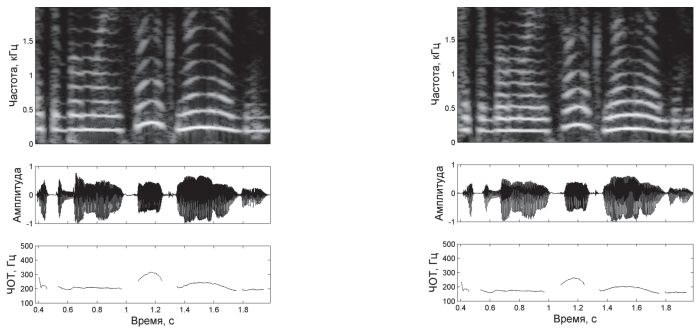
а б

Рис. 12. Эффект повышения частоты основного тона;
а – входной сигнал, б – обработанный сигнал

Второй эффект заключался в понижении основного тона. Целевая частота основного тона определялась из исходной как:

$$\bar{F}_0(n) = \frac{F_0(n)}{k}, \qquad k > 1,$$

где k – коэффициент, на который происходит понижение тона. Высокая натуральность синтезированного голоса достигается, если коэффициент k не превышает 2. На рисунке 13 показан пример повышения основного тона.



а б

Рис. 13. Эффект понижения частоты основного тона;
а – входной сигнал, б – обработанный сигнал

Третий эффект ‘*cycle*’ выполняет периодическое повышение тона. Профиль согласно которому происходит повышение задается функцией $f_{profile}(n)$, которая определена на конечном интервале от 0 до $N-1$. Пример такой функции показан на рисунке 14.

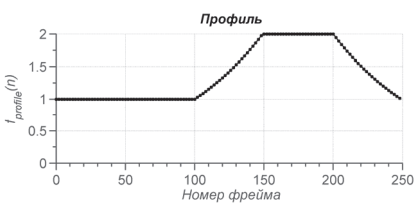
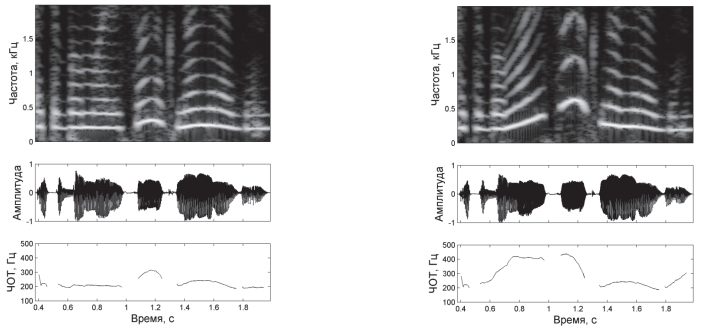


Рис. 14. Функция-профиль для изменения частоты основного тона

При реализации эффекта ‘*cycle*’ целевая частота основного тона определяется следующим образом:

$$\bar{F}_0(n) = F_0(n) \cdot f_{profile}(n \bmod N).$$

На рисунке 15 показан результат применения эффекта ‘*cycle*’ к речевому сигналу.



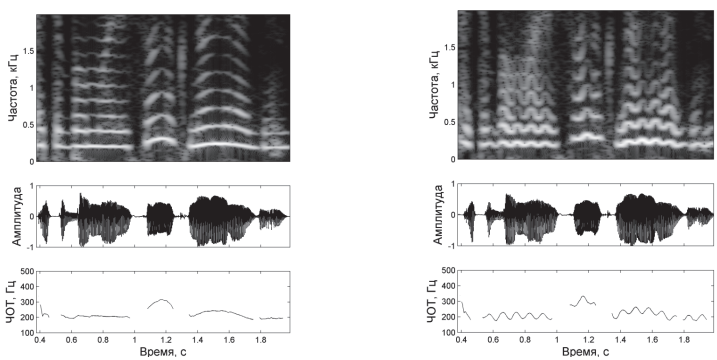
а б

Рис. 15. Эффект ‘*cycle*’;
а – входной сигнал, б – обработанный сигнал

Четвертый эффект ‘*sinus*’ позволяет придать голосу дрожание, которое иногда появляется у человека в момент волнения. Для достижения этого эффекта частота основного тона сигнала преобразуется следующим образом:

$$\bar{F}_0(n) = F_0(n) + 20\cos\left(2\pi n\frac{12}{F_s}\right),$$

где F_s – частота дискретизации сигнала. На рисунке 16 показан результат применения эффекта ‘*sinus*’ к речевому сигналу.



а б

Рис. 16. Эффект ‘*sinus*’;
а – входной сигнал, б – обработанный сигнал

5.2. Речевая база и оценка качества обработки. Тестирование системы изменения основного тона речевого сигнала выполнялось на речевой базе, начитанной 9 дикторами (4 мужских голоса и 5 женских). Длительность записи каждого диктора приблизительно составляла 1 мин 10 сек. Чтобы оценить качество работы системы в телекоммуникационной системе на ряду с исходными сигналами обрабатывались сигналы декодированные кодеками GSM и G.711.

Образцы сигналов, обработанные четырьмя различными эффектами, прослушивались группой экспертов. Задачей экспертов было оценить натуральность речи по 5 бальной шкале ("5" – натуральная речь; "4" – речь звучит натурально, заметны отдельные артефакты; "3" – синтезированная речь, имеются артефакты; "2" – ненатуральное звучание, отдельные слова не воспринимаются; "1" – полная потеря разборчивости). В таблице 1 приводятся средние оценки, полученные в результате эксперимента.

Таблица 1. Оценка натуральности образцов речи, обработанных в системе

Эффект / Образец	Без обработки	Повышение тона $k = 1.2$	Понижение тона $k = 1.2$	Эффект *cycle*	Эффект *sinus*
Широкополосная речь	5	4.80	4.75	4.30	3.95
Речь после кодека GSM	4.25	4.15	4.20	3.85	3.70
Речь после кодека G.711	4.15	4.05	4.15	3.85	3.75
Речь после кодеков G.711 и GSM	4.10	4.00	4.10	3.80	3.60

На вход системы подавались сигналы трех различных категорий: широкополосная речь, узкополосная речь и узкополосная речь, декодированная после сжатия кодеками GSM и G.711. Таким образом, для каждой категории сигнала и каждого эффекта изменения тона было синтезировано 9 различных образцов по числу дикторов. Каждый из полученных образцов оценивался группой экспертов, а в таблицу заносились только средние значения оценок.

Очевидно, что натуральность речи будет тем выше, чем меньше изменений вносится в контур частоты основного тона. Это подтверждается и экспериментальными данными. Эффекты понижения и повышения вносят наименьшие изменения и по этой причине они получили более высокую оценку. Эффект ‘*cycle*’ имеет временные интервалы, на которых динамика изменения контура частоты основного тона остается неизменной, поэтому натуральность этого эффекта выше чем

у эффекта '*sinus*', который вносит систематическое изменение в натуральную форму контура.

6. Заключение. В работе предложено решение задачи изменения основного тона речи на основе гибридной модели речевого сигнала. Обработка вокализованной составляющей сигнала выполняется в измененном масштабе времени, что позволяет выполнять точное разделение гармоник и оценивать их мгновенные параметры при помощи узкополосной фильтрации. Для снижения вычислительной сложности используется полифазная реализация ДПФ-модулированного банка фильтров. Полученные результаты моделирования свидетельствуют о применимости предложенного решения в широкополосных и узкополосных каналах связи.

Литература

1. *Flanagan J.L., Golden R. M.* Phase vocoder // Bell System Technical Journal, 1966. vol. 45, pp. 1493-1509.
2. *Levine S., Smith J.* A sines+transients+noise audio representation for data compression and time/pitch scale modifications // Signal processing: proceedings of 105th AES convention, San Francisco, USA, San Francisco, Preprint 1998. № 4781. 21 p.
3. *Serra X.* A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition: PhD thesis // Stanford, 1989. 178 p.
4. *Kawahara H., Takahashi T., Morise M., Banno H.* Development of exploratory research tools based on TANDEM-STRAIGHT // Proc. APSIPA, Sapporo, Japan, Oct. 2009.
5. *Kawahara H., Morise M.* Analysis and synthesis of strong vocal expressions: extension and application of audio texture features to singing voice // Proc. ICASSP'2012, Kyoto, Japan, March 2012. pp. 5389–5392.
6. *Erro D., Sainz I., Navas E., Hernaez I.* Improved HNM-based vocoder for statistical synthesizers // Proc. INTERSPEECH, Florence, Italy, Aug. 2011.
7. *Painter T., Spanias A.* Sinusoidal analysis-synthesis of audio using perceptual criteria // EURASIP Journal on Applied Signal Processing. 2003. № 1. pp. 15-20.
8. *Degottlex G., Stylianou Y.* A full-band adaptive harmonic representation of speech // Proc. INTERSPEECH, Portland, Oregon, USA, Sep. 2012.
9. *Azarov E., Vashkevich M., Petrovsky A.* Instantaneous pitch estimation based on RAPT framework // Proc. EUSIPCO, Bucharest, Romania, Aug. 2012, pp. 2787-2791.
10. *Азаров И.С., Вашкевич М.И., Петровский А.А.* Алгоритм оценки мгновенной частоты основного тона речевого сигнала // Цифровая обработка сигналов. Москва: 2012. №4. С. 49-57.
11. *Talkin D.* A Robust Algorithm for Pitch Tracking (RAPT) // Speech Coding & Synthesis, W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.
12. *Vaidynathan P.P.* Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial // Processing of the IEEE. January, 1990. vol. 78, no 1. pp. 56–93.

References

1. Flanagan J.L., Golden R. M. Phase vocoder. Bell System Technical Journal, 1966. vol. 45, pp. 1493-1509.

2. Levine S., Smith J. A sines+transients+noise audio representation for data compression and time/pitch scale modifications. Signal processing: proceedings of 105th AES convention, San Francisco, USA, San Francisco, Preprint 1998. no. 4781. 21 p.
3. Serra X. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition: PhD thesis. Stanford, 1989. 178 p.
4. Kawahara H., Takahashi T., Morise M., Banno H. Development of exploratory research tools based on TANDEM-STRAIGHT. Proc. APSIPA, Sapporo, Japan, Oct. 2009.
5. Kawahara H., Morise M. Analysis and synthesis of strong vocal expressions: extension and application of audio texture features to singing voice. Proc. ICASSP'2012, Kyoto, Japan, March 2012. pp. 5389–5392.
6. Erro D., Sainz I., Navas E., Hernaez I. Improved HNM-based vocoder for statistical synthesizers. Proc. INTERSPEECH, Florence, Italy, Aug. 2011.
7. Painter T., Spanias A. Sinusoidal analysis-synthesis of audio using perceptual criteria. EURASIP Journal on Applied Signal Processing. 2003. no. 1. pp. 15-20.
8. Degottlex G., Stylianou Y. A full-band adaptive harmonic representation of speech. Proc. INTERSPEECH, Portland, Oregon, USA, Sep. 2012.
9. Azarov E., Vashkevich M., Petrovsky A. Instantaneous pitch estimation based on RAPT framework. Proc. EUSIPCO, Bucharest, Romania, Aug. 2012, pp. 2787-2791.
10. Azarov E., Vashkevich M., Petrovsky A. [An algorithm for instantaneous pitch estimation of speech]. *Cifrovaja obrabotka signalov – Digital Speech Processing*. Moscow: 2012. no. 4, pp. 49-57. (In Russ.).
11. Talkin D. A Robust Algorithm for Pitch Tracking (RAPT). Speech Coding & Synthesis, W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.
12. Vaidynathan P.P. Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial. Processing of the IEEE. January, 1990. vol. 78, no 1. pp. 56–93.

Азаров Илья Сергеевич — к-т тех. наук., доцент кафедры электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка речевых сигналов. Число научных публикаций — 42. azarov@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-8805.

Azarov Elias — Ph.D., associate professor of computer engineering department, BSUIR. Research interests: digital speech processing. The number of publications — 42. azarov@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-8805.

Вашкевич Максим Иосифович — к-т тех. наук., доцент кафедры электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка сигналов. Число научных публикаций — 36. vashkevich@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-8805.

Vashkevich Maxim — Ph.D., associate professor of computer engineering department, BSUIR. Research interests: digital signal processing. The number of publications — 36. vashkevich@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-8805.

Лихачев Денис Сергеевич — к-т тех. наук, доцент, доцент кафедры электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка сигна-

лов. Число научных публикаций — 40. likhachov@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-8805.

Likhachov Denis — Ph.D., associate professor; associate professor of computer engineering department, BSUIR. Research interests: digital speech processing. The number of publications — 40. vashkevich@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-8805.

Петровский Александр Александрович — д-р тех. наук, профессор, заведующий кафедрой электронных вычислительных средств БГУИР. Область научных интересов: цифровая обработка сигналов. Число научных публикаций — более 600. palex@bsuir.by, www.bsuir.by; БГУИР, ул. П. Бровки 6, г. Минск, 220013, РБ; р.т. +375 (17) 293-2340.

Petrovsky Alexander — Ph.D., Dr. Sci., professor, head of computer engineering department, BSUIR. Research interests: digital speech processing. The number of publications — more than 600. palex@bsuir.by, www.bsuir.by; BSUIR, 6, P.Brovky str., 220013, Minsk, RB; office phone +375 (17) 293-2340.

РЕФЕРАТ

Азаров И.С., Вашкевич М.С., Лихачев Д.С., Петровский А.А. **Изменение частоты основного тона речевого сигнала на основе гармонической модели с нестационарными параметрами.**

В статье предлагается решение задачи изменения частоты основного тона речевого сигнала. Необходимость решения данной задачи возникает во многих речевых приложениях таких как конверсия голоса, коррекция акцента, обеспечение конфиденциальности диктора и др.

Разработанная схема обработки вокализованной части речевого сигнала основывается на гармонической модели с нестационарными (изменяющимися в каждый момент времени) параметрами. Оценка параметров модели выполняется с учетом контура мгновенной частоты основного тона. Входной речевой сигнал масштабируется во времени для того, чтобы обеспечить его стационарность, а затем выполняется узкополосная фильтрация, разделяющая гармоники основного тона. В результате фильтрации формируются аналитические (комплексные) сигналы, которые описываются при помощи параметров синусоидальной модели. На основе анализа смежных значений мгновенной частоты определяется степень вокализации каждой гармоники. Затем выполняется синтез квазипериодического сигнала по полученным параметрам, который вычитается из исходного речевого сигнала. В результате выполняется разделение речевого сигнала на вокализованную и невокализованную части. Одновременно синтезируется квазипериодический сигнал с целевой частотой основного тона, который складывается с выделенной невокализованной частью.

На основании субъективной оценки результатов показано, что разработанный способ обеспечивает высокую натуральность и разборчивость синтезированной речи и может применяться как в широкополосных так и в узкополосных каналах связи с различными стандартами кодирования (в том числе с кодеками G.711 и GSM).

SUMMARY

Azarov E., Vashkevich M., Likhachov D., Petrovsky A. **Pitch modification of speech signal using harmonic model with time-varying parameters.**

The paper presents a solution to the problem of pitch modification of speech. The problem occurs in different speech processing applications such as voice conversion, accent correction, hiding speaker's personality and other.

Developed processing scheme for voiced part of speech is based on the harmonic model with nonstationary (time-varying) parameters. Parameters of the model are estimated synchronously with intstantaneous contour of pitch. Input speech is warped in time domain in order to ensure stationarity of pitch and then narrow-band filtering is applied which separates individual pitch harmonics. The filtering results in analytical (complex) subband signals which can be represented by means of sinusoidal modeling. Vocalization degree of each harmonic is extracted by analysis of instantaneous frequency values of adjacent frames. Vocalized part of the signal is synthesized using extracted sinusoidal parameters and subtracted from the source signal. That results in deterministic/stochastic decomposition of the signal. The new vocalized part of the signal with target pitch values is synthesized and added to the extracted stochastic part.

Using subjective listening tests it is shown that developed system provides high naturalness and intelligibility of reconstructed speech and can be applied to wideband and narrowband communication channels with various coding standards (including G.711 and GSM).