ИНФОРМАТИКА

УДК 004.934 + 004.4'277

ОБРАБОТКА РЕЧЕВЫХ СИГНАЛОВ В ПРИЛОЖЕНИЯХ МУЛЬТИМЕДИА НА ОСНОВЕ ПЕРИОДИЧЕСКОЙ МОДЕЛИ С НЕСТАЦИОНАРНЫМИ ПАРАМЕТРАМИ

А.А. ПЕТРОВСКИЙ, И.С. АЗАРОВ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220027, Беларусь

Поступила в редакцию 8 января 2014

Рассматриваются методы нестационарной параметризации речевых сигналов, позволяющие выполнять анализ, обработку и синтез речи в приложениях мультимедиа. Формулируются основные теоретические положения и рассматриваются вопросы практической реализации. Приводятся результаты применения методов к задачам оценки основного тона и изменения просодических характеристик речевого сигнала.

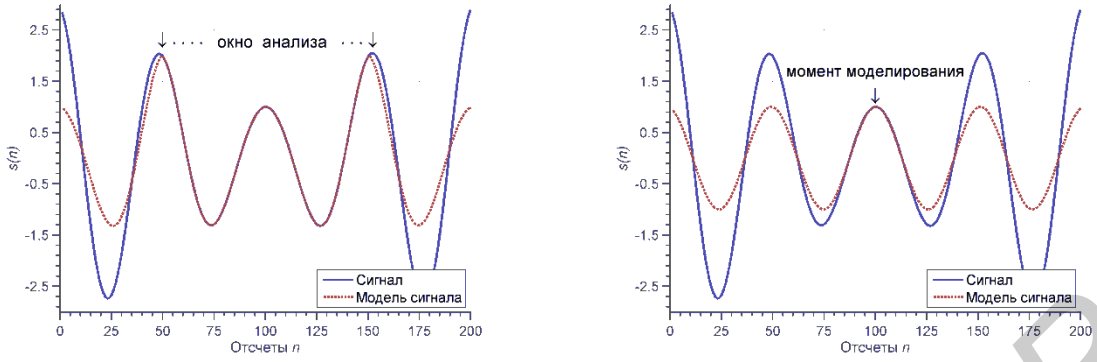
Ключевые слова: обработка речевых сигналов, оценка мгновенной частоты основного тона, оценка гармонических параметров, параметрический анализ и синтез речи.

Введение

Различные способы параметрического представления речевых сигналов используются при решении таких сложных задач как создание речевых интерфейсов, распознавание речи, синтез речи по тексту, конверсия голоса, шумоподавление, повышение разборчивости и субъективного качества речевых сигналов, коррекция акцента, синтез обучающих речевых сообщений и т. д. В самом общем виде модель речевого сигнала обычно содержит две основные составляющие: спектральную огибающую и сигнал возбуждения [1–4]. Спектральная огибающая определяет фонетику произносимого звука и характеризует состояние речевого тракта, в то время как сигнал возбуждения характеризует состояние голосовых связок и высоту (интонацию) вокализованных звуков. Каждая из этих составляющих выделяется при помощи речевого анализатора и описывается своим набором параметров.

Процесс обработки речевого сигнала обычно включает анализ (определение параметров модели), модификацию (изменение параметров модели в зависимости от цели приложения) и синтез (формирование нового сигнала из измененных параметров модели). Учитывая, что различные части речи имеют разную природу речеобразования (вокализованную и невокализованную) используется гибридный подход к описанию сигнала, который заключается в том, что сигнал разделяется на две составляющие: квазипериодическую (детерминистскую) и непериодическую (стохастическую). Каждая из этих составляющих моделируется отдельно.

По способу параметризации речевые модели можно разделить на две группы «стационарные» и «нестационарные». При «стационарном» моделировании сигнал на протяжении некоторого интервала наблюдения представляется постоянными параметрами. В качестве средств анализа могут использоваться различные методы, в числе которых преобразование Фурье, метод Прони и линейное предсказание. При «нестационарном» моделировании сигнал в каждый момент времени представляется в виде отдельного набора параметров (рис. 1).



а *б*

Рис. 1 – Моделирование периодического сигнала:
а – стационарное моделирование; б – нестационарное моделирование

Наиболее традиционными способами оценки мгновенных (т.е. относящихся к данному моменту времени) параметров являются преобразование Гильберта и алгоритм разделения энергии ESA (Energy Separation Algorithm). Оба эти подхода требуют декомпозиции сигнала на однокомпонентные периодические функции. Существует метод моделирования речевого сигнала путем разделения его на узкополосные комплексные составляющие (аналитические сигналы) при помощи фильтрации. Каждая из составляющих соответствует одной гармонике основного тона и описывается одной комплексной экспонентой, имеющей мгновенную амплитуду, фазу и частоту. Поскольку вокализованная речь состоит из квазипериодических компонент с изменяющимися параметрами, при фильтрации возникают сложности, связанные с необходимостью применять фильтры с изменяющимися характеристиками: полоса пропускания должна меняться в соответствии с контуром частоты основного тона. Для анализа речевых сигналов были предложены специальные частотно-временные преобразования, позволяющие производить адекватную оценку компонент с сильной частотной модуляцией, такие как Фан-Чирп и гармоническое [4–6] преобразования. Однако гармонические амплитуды принимаются постоянными на протяжении всего фрейма анализа, что существенно ограничивает точность.

Параметрическое представление речи подразумевает использование частоты основного тона (ЧОТ) в качестве параметра модели. Выбор определенного алгоритма для оценки частоты основного тона зависит от целевого приложения и всегда представляет собой некоторый компромисс между частотно-временным разрешением, устойчивостью к ошибкам, алгоритмической задержкой и вычислительной сложностью. Точность оценки ЧОТ определяет насколько хорошо можно разделить сигнал на детерминистическую и стохастическую составляющие, от нее зависит также число разделяемых гармоник, которые можно описать отдельными наборами параметров. Точность определяется двумя основными характеристиками: 1 – временное разрешение, т. е. как быстро алгоритм оценки реагирует на изменения частоты, 2 – частотное разрешение, т. е. насколько малые изменения частоты алгоритм может определить. Обе характеристики чувствительны к модуляциям основного тона и степени зашумленности сигнала (интенсивности шума как фонового так и обусловленного смешанным возбуждением речевого тракта). В настоящее время предложено большое число разнообразных алгоритмов оценки основного тона, наиболее популярными из них являются RAPT, YIN и SWIPE’. Популярность данных алгоритмов обусловлена хорошей функциональностью, низким процентом грубых ошибок и наличием свободно распространяемых версий их реализаций. Тем не менее, возможность этих алгоритмов оценивать мгновенную частоту существенно ограничена. Ограничение обусловлено периодической (стационарной) моделью сигнала, лежащей в их основе, которая подразумевает точное повторение периода основного тона и не допускает его изменения на протяжении анализируемого фрейма. При появлении модуляций (изменений частоты основного тона) точность оценок существенно снижается. В последнее время было предложено несколько оригинальных методов для оценки мгновенной ЧОТ, однако, они не имеют свободных

программных реализаций, доступных для использования и тестирования.

В настоящей работе приводится оригинальный способ оценки нестационарной частоты основного тона, основанный на специальной функции оценки периодичности, а также способ оценки нестационарных гармонических параметров, учитывающий модуляции ЧОТ. Реализация данных методов позволяет выполнять моделирование речевых сигналов в различных приложениях мультимедиа.

Оценка нестационарных параметров вокализованной речи

Задача оценки нестационарных параметров квазипериодических сигналов сводится к определению амплитуды $A_k(t)$, частоты $f_k(t)$ и фазы $\varphi_k(t)$ каждой его составляющей $k = 1,2,...,K$ как функций, зависящих от времени. Оценка параметров должна проводиться, исходя из предположения, что компоненты могут быть частотно-модулированы и их параметры изменяются в каждый момент времени. Если предположить, что составляющие сигнала занимают неперекрывающиеся частотные диапазоны (данное предположение справедливо для вокализованной речи) то их можно разделить при помощи узкополосной фильтрации [4–10]. Для синтеза соответствующих цифровых фильтров можно воспользоваться оконным методом. Пусть F_1 и F_2 – нормированные частоты из диапазона $[0,\pi]$, определяющие соответственно нижнюю и верхнюю границы полосы пропускания, причем $F_2 > F_1$, тогда непрерывная импульсная характеристика искомого фильтра $h(t)$ определяется следующим выражением:

$$h_{F_1,F_2}(t) = \frac{1}{\pi}\int_0^{F_2} e^{-j\omega t}d\omega - \frac{1}{\pi}\int_0^{F_1} e^{-j\omega t}d\omega = \frac{e^{-j\omega t}}{-jt\pi}\Bigg|_0^{F_2} - \frac{e^{-j\omega t}}{-jt\pi}\Bigg|_0^{F_1} = \frac{e^{-jF_1 t} - e^{-jF_2 t}}{jt\pi}.$$

Выразим импульсную характеристику через середину полосы пропускания F_C и половину ширины полосы F_Δ, используя подстановку $F_1 = F_C - F_\Delta$ и $F_2 = F_C + F_\Delta$:

$$h_{F_1,F_2}(t) = \frac{e^{-jF_C t}e^{jF_\Delta t} - e^{-jF_C t}e^{-jF_\Delta t}}{jt\pi} = \frac{e^{-jF_C t}\left(e^{jF_\Delta t} - e^{-jF_\Delta t}\right)}{jt\pi} = 2\frac{\sin(F_\Delta t)}{t\pi}e^{-jF_C t}.$$

Сигнал на выходе фильтра представляет собой АМ и ЧМ функцию косинуса с ограниченным частотным диапазоном:

$$S_{F_1,F_2}(t) = s(t) * h_{F_1,F_2}(t) = A_{F_1,F_2}(t)\cos(\varphi_{F_1,F_2}(t))$$

с мгновенной амплитудой $A_{F_1,F_2}(t)$, фазой $\varphi_{F_1,F_2}(t)$ и частотой $f_{F_1,F_2}(t)$, которые могут быть определены по соответствующим формулам:

$$A_{F_1,F_2}(t) = \sqrt{R^2(t) + I^2(t)},\ \varphi_{F_1,F_2}(t) = \arctan\left(\frac{-I(t)}{R(t)}\right),\ f_{F_1,F_2}(t) = \varphi'_{F_1,F_2}(t),$$

где $R(t)$ и $I(t)$ – действительная и мнимая части комплексного сигнала $S_{F_1,F_2}(t)$ соответственно. Для получения импульсной характеристики цифрового фильтра конечной длины следует использовать некоторую оконную функцию $w(t)$:

$$h_{F_1,F_2}(t) = 2\frac{\sin(F_\Delta t)}{t\pi}w(t)e^{-jF_C t}.$$

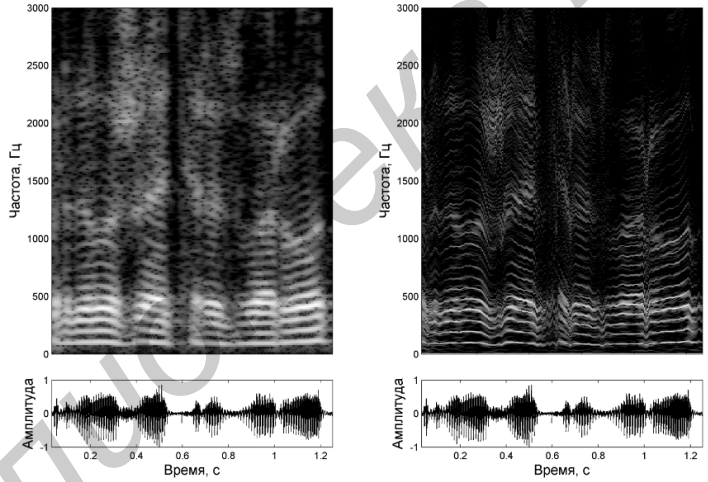
Заметим, что $h_{F_1,F_2}(t)$ представляет собой произведение импульсной характеристики идеального фильтра низких частот, оконной функции и комплексной экспоненты, выполняющей частотный сдвиг на заданную частоту F_C. Параметры сигнала в любой заданный момент времени t_0 можно определить при помощи следующего выражения:

$$S(F_\Delta, F_C, t_0) = \int_{-\infty}^{+\infty} 2\frac{\sin(F_\Delta t)}{t\pi} w(t)s(t_0 - t)e^{-jF_C t}dt.$$

Для точной оценки синусоидальных параметров компонент с сильной частотной модуляцией должен быть использован частотно-модулированный фильтр, импульсная характеристика которого модулируется в соответствии с частотным контуром анализируемой компоненты. Приблизительные частотные траектории гармоник могут быть получены из контура частоты основного тона. Непрерывность импульсной характеристики синтезированного стационарного фильтра позволяет достаточно просто адаптировать его к частотным модуляциям. Рассматривая центральную частоту полосы пропускания как функцию от времени $F_C(t)$, можно применить фильтрацию с частотным масштабированием:

$$S(F_\Delta, F_C(t), t_0) = \int_{-\infty}^{+\infty} 2\frac{\sin(F_\Delta t)}{t\pi} w(t)s(t_0 - t)e^{-j\varphi_C(t,t_0)}dt,$$

где $\varphi_C(t, t_0) = \int_{t_0}^{t} F_C(t) - F_C(t_0)dt$. Используя полученный фильтр, анализируемая частотно-модулированная компонента может быть выделена из узкой масштабированной полосы пропускания, что позволяет применять мгновенный гармонический анализ к гармоникам высокого порядка. Чем выше номер гармоники, тем больше изменение ее частоты и импульсная характеристика ЧМ-фильтра изменяется соответствующим образом – рис. 2.



а *б*

Рис. 2. Оценка спектральных компонент речевого сигнала (окно анализа 250 мс):
а – преобразование Фурье (окно анализа 64 мс); *б* – узкополосная фильтрация, согласованная с непрерывным контуром основного тона (окно анализа 250 мс)

Оценка мгновенной частоты основного тона

Одним из традиционных способов генерации кандидатов периода основного тона является автокорреляционная функция. Пусть $s(m)$ – анализируемый дискретный сигнал, z – величина шага в отсчетах и n – размер окна, тогда автокорреляционная функция $R(x,k)$ для K отсчетов, задержки k и анализируемого фрейма x определяется как

$$R(x,k) = \sum_{i=m}^{m+n-k-1} s(i)s(i+k),\ k = 0, K-1; m = xz; x = 0, M-1.$$

Благодаря относительной устойчивости к шуму автокорреляционная функция с успехом используется во многих алгоритмах оценки ЧОТ. Тем не менее, она имеет ряд недостатков,

которые ограничивают ее использование в качестве функции генератора кандидатов периода. Основным из недостатков является необходимость использовать продолжительные окна анализа для того чтобы оценить периодичность сигнала во всем интересующем диапазоне. В результате резкие изменения ЧОТ приводят к потере четких пиков $R(x,k)$ в точке, соответствующей действительному периоду. Другим недостатком является неодинаковое число отсчетов, участвующих в оценке $R(x,k)$ для разных задержек *k*. Это приводит к тому, что устойчивость автокорреляционной функции к шумам так же зависит от задержки и если для больших значений *k* окно анализа достаточно по длине, то для малых оно избыточно.

Периодичность фрагмента сигнала удобно определять при помощи нормированной кросс-корреляционной функции (НККФ) $\phi(x,k)$, в которой недостатки автокорреляционной функции менее выражены. НККФ определяется как

$$\phi(x,k)=\frac{\sum_{i=m}^{m+n-1}s(i)s(i+k)}{\sqrt{e_m e_{m+k}}},\ k=0,K-1;m=xz;x=0,M-1,\text{ где } e_i=\sum_{l=i}^{i+n-1}s_l^2.$$

Следует отметить, что значения $\phi(x,k)$ находятся в диапазоне от –1 до +1, причем функция приближается к верхнему пределу для задержек, кратных действительному периоду основного тона вне зависимости от амплитуды анализируемого сигнала. Допустимый диапазон периода основного тона не зависит от продолжительности окна анализа. Если анализируемый сигнал является белым шумом, то $\phi(x,k)$ будет приближаться к нулю для всех $k>0$ при увеличении длины окна анализа.

В предлагаемом алгоритме оценки ЧОТ функция $\phi(x,k)$ оценивается при помощи нестационарных параметров сигнала. Параметрическое представление каждого отсчета *s(m),* определяемое квазипериодической моделью, может быть использовано для вычисления мгновенной автокорреляционной функции $R_{inst}(m,k)$, используя теорему Винера-Хинчина [3]:

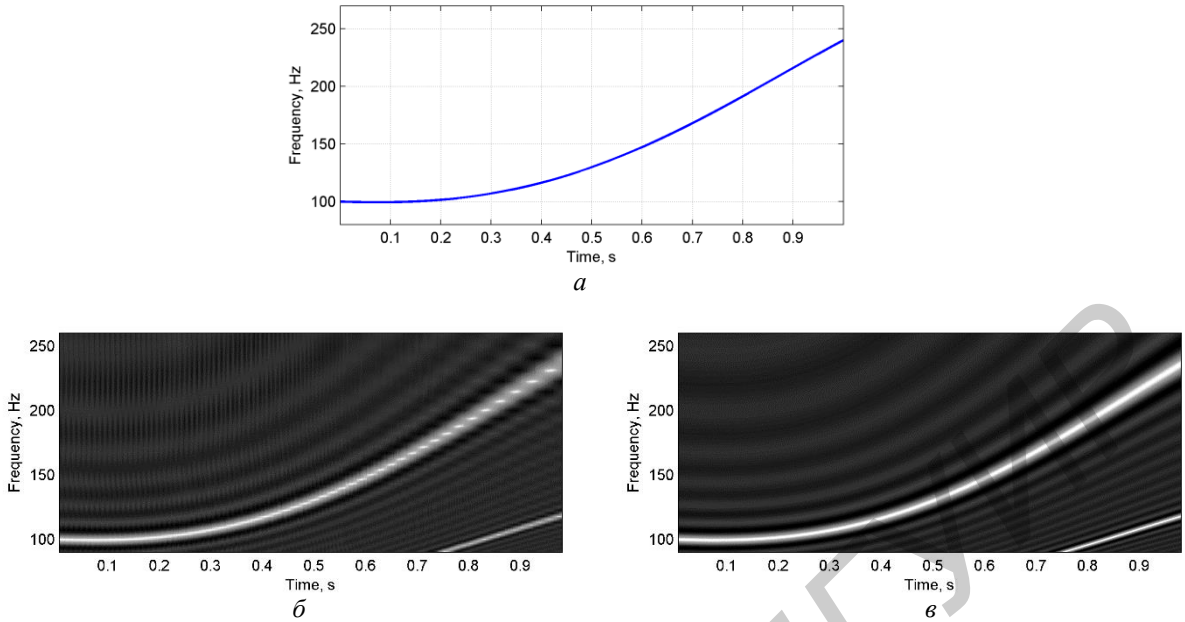
$$R_{inst}(m,k)=\frac{1}{2}\sum_{p=1}^{P}A_p^2(m)\cos(F_p(m)k)$$

где $A_p^2(m)$ – нестационарная амплитуда, $F_p^2(m)$ – мгновенная частота, P – число квазипериодических составляющих.

$R_{inst}(m,k)$ соответствует автокорреляционной функции, вычисленной для периодического сигнала бесконечной длины с постоянными значениями A_p и F_p. Поскольку окно анализа в данном случае бесконечно, то не будет разницы между нормированной автокорреляционной функцией и НККФ. Следовательно, НККФ можно оценить через мгновенные параметры синусоидальной модели следующим образом:

$$\phi_{inst}(m,k)=\frac{\sum_{p=1}^{P}A_p^2(m)\cos(F_p(m)k)}{\sum_{p=1}^{P}A_p^2(m)}$$

Особенностью этой функции является то, что в отличие от НККФ, задержка *k* не обязательно должна быть целой и, таким образом, можно получить оценку периодичности для любого вещественного периода. Вторым важным отличием является то, что предлагаемая функция нечувствительна к любым изменениям частоты основного тона в окрестности отсчета *m* при условии, что полученные гармонические параметры являются достаточно точными. На рис. 3 показано, что для частотно-модулированного сигнала традиционная НККФ подвержена «эффекту ступенек», в то время как НККФ на основе нестационарной модели формирует непрерывный контур кандидатов искомого периода основного тона.



а *б* *в*

Рис. 3. Функции генерации кандидатов искомого периода основного тона:
а – действительный контур частоты основного тона; *б* – НККФ; *в* – НККФ на основе нестационарных гармонических параметров

Учитывая то, что после первоначальной оценки основного тона каждый узкополосный аналитический сигнал соответствует одной гармонике основного тона, уточнение частоты основного тона может быть выполнено при помощи взвешенного среднего:

$$F_0(m) = \sum_{p=1}^{P} \frac{F_p(m) A_p(m)}{p \sum_{j=1}^{P} A_j(m)}$$

Оценка спектральной огибающей

Традиционно для оценки огибающей спектра речевого сигнала используют кепстральный анализ либо линейное предсказание. Причем в линейном предсказании применяются предсказатели ограниченных порядков, поскольку с увеличением числа коэффициентов частотная характеристика фильтра-предсказателя вместо спектральной огибающей сигнала начинает описывать поведение отдельных гармоник. Используя полученные нестационарные гармонические параметры возможно точно описать спектральную огибающую, определяемую параметрами гармоник основного тона, при помощи фильтра-предсказателя высокого порядка. Причем, поскольку выполняется непосредственная конверсия одних параметров в другие, временное разрешение оценки огибающей будет очень высоким.

Показано, что коэффициенты фильтра-предсказателя порядка *p* могут быть получены при помощи системы линейных уравнений [4]:

$$\sum_{i=1}^{p} a_i q(|i-j|) = -q(j),$$

где $j = 1,2,\ldots,p$ и $q(l) = \sum_{k=1}^{K} A_k(n) \cos(f_k(n) l)$, $(l \geq 0)$.

Если целевую спектральную огибающую, рассматривать как непрерывную функцию от частоты $A(\omega)$, заданную на интервале $[0,\pi]$, то элементы матрицы системы преобразования $q(l)$ могут быть вычислены в виде интеграла:

$$q(l) = \int_0^{\pi} A(\omega)\cos(\omega l)d\omega.$$

Если функция $A(\omega)$ содержит точки разрыва $\omega_d = (\omega_1, \omega_2, ..., \omega_I)$, тогда

$$q(l) = \sum_{i=1}^{I+1} \int_{\overline{\omega}_{d,i}}^{\overline{\omega}_{d,i+1}} A(\omega)\cos(\omega l)d\omega.$$

где $\overline{\omega}_d = (0, \omega_1, \omega_2, ..., \omega_I, \pi)$.

Например, если определить функцию амплитудной огибающей в виде амплитудно-частотной характеристики полосового фильтра

$$A(\omega) = \begin{cases} 1, & F_1 \le \omega \le F_2 \\ 0, & \omega \le F_1, \omega \ge F_2 \end{cases}, \ 0 \le F_1 \le F_2 \le \pi,$$

то выражение примет вид

$$q(l) = \begin{cases} \sin(F_2 l)/l, & l \ne 0 \\ F_2 - F_1, & l = 0 \end{cases}$$

и в результате решения соответствующей системы обратный фильтр-предсказатель будет представлять собой полосовой фильтр с полосой пропускания $F_1 \le \omega \le F_2$. Ниже представлены амплитудно-частотные характеристики двух полосовых фильтров – один синтезирован при помощи оконного метода (использовалось окно Хэмминга), а второй синтезирован при помощи метода, описанного выше. Фильтры синтезированы с одной полосой пропускания $0{,}2\pi \le \omega \le 0{,}3\pi$ и одинаковым числом коэффициентов – рис. 4.

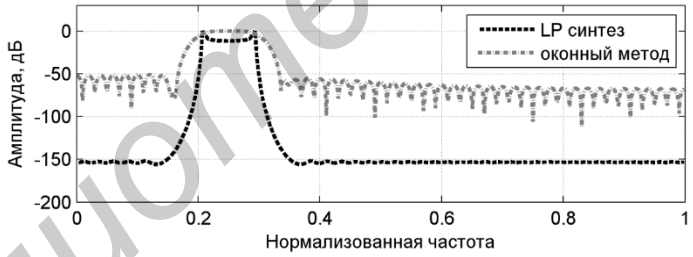


Рис. 4. Амплитудно-частотные характеристики полосовых фильтров, синтезированных при помощи линейного предсказания и оконного метода

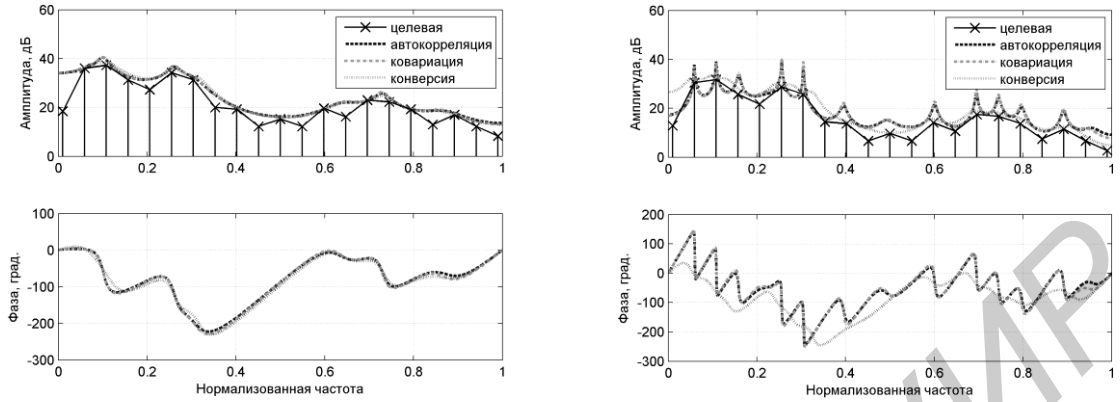
Непрерывная амплитудная огибающая спектра может быть получена из векторов амплитудных и частотных значений путем линейной интерполяции. Каждый сегмент огибающей $f_i \le \omega \le f_{i+1}, \quad 1 \le i \le K-1$ описывается линейным уравнением прямой $A(\omega) = b_i\omega + c_i$. Параметры b_i и c_i вычисляются из смежных значений амплитуды и частоты. Элементы системы линейных уравнений принимают вид

$$q(l) = \sum_{i=1}^{K-1} D(l,i)$$

где

$$D(l,i) = \begin{cases} b/l^2\left[\cos(f_{i+1}l) + f_{i+1}l\sin(f_{i+1}l)\right] + c/l(\sin(f_{i+1}l) - \sin(f_i l)) - \\ -b/l^2\left[\cos(f_i l) + f_i l\sin(f_i l)\right] & l \ne 0 \\ b/2\left(f_{i+1}^2 - f_i^2\right) + c\left(f_{i+1} - f_i\right) & l = 0 \end{cases}$$

Ниже показано, как описанный способ оценки спектральной огибающей соотносится с другими методами линейного предсказания на примере анализа полигармонического сигнала с известными параметрами (рис. 5).



а *б*

Рис. 5. Оценка огибающей гармонического сигнала при помощилинейного предсказания: *а* – 14 коэффициентов предсказания; *б* – 30 коэффициентов предсказания

В приведенном примере для получения коэффициентов линейного предсказания автокорреляционным и ковариационным методами использовался синтезированный полигармонический сигнал из 1024 отсчетов. Для синтеза сигнала использовались заданные векторы значений частоты и амплитуды целевой огибающей. Из рис. 5 видно, что все методы показывают близкие оценочные огибающие в случае 14-ти коэффициентов предсказания, однако при увеличении числа коэффициентов описанная выше техника обеспечивает намного более точный результат.

Практическая реализация системы моделирования

Процедура обработки речи с использованием нестационарной параметрической модели состоит из последовательности шагов, схематически показанных на рис. 6. Речевой сигнал представляется в виде наборов параметров, относящихся к определенным моментам времени с постоянным шагом в несколько миллисекунд. В каждый момент сигнал классифицируется как вокализованный либо невокализованный. Классификация выполняется при помощи специального детектора, анализирующего форму спектральных огибающих. Параметрическое представление невокализованных участков речи выполняется при помощи псевдослучайной последовательности (белого шума), проходящей через фильтр, аппроксимирующий заданную спектральную плотность мощности.

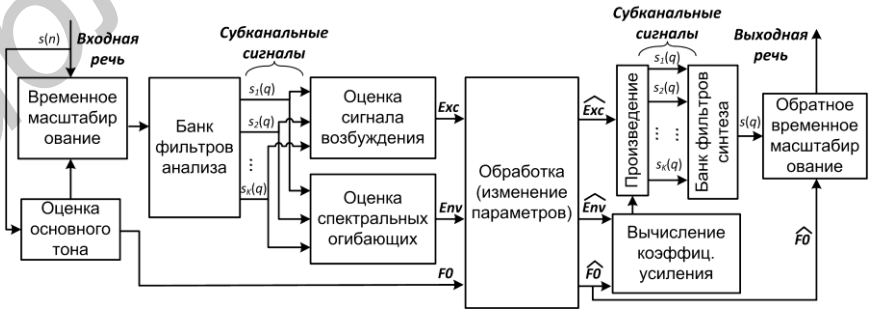


Рис. 6. Общая схема обработки речевого сигнала

Для оценки мгновенной частоты основного тона вокализованных участков речи используется модифицированный алгоритм слежения, устойчивый к ошибкам [3]. Алгоритм использует мгновенные гармонические параметры для вычисления НККФ как было показано выше, что позволяет получать устойчивые к модуляциям мгновенные оценки частоты основного тона.

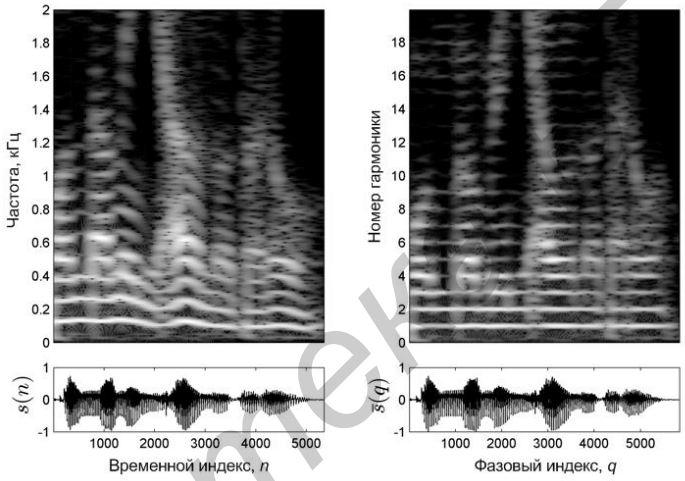
Процедура временного масштабирования выполняется путем вычисления значений сигнала $s(n)$ в новые моменты времени $m(q)$ таким образом, чтобы на каждый период основного тона приходилось равное число отсчетов N_{f0}. Для каждого отсчета исходного сигнала $s(n)$ формируется фазовая метка $\phi(n)$, используя мгновенные значения основного тона $f_0(n)$:

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

Новые моменты времени $m(q)$ вычисляются как

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

где q – индекс отсчетов сигнала в измененном масштабе времени $\bar{s}(q)$. Полученный в результате сигнал имеет постоянную частоту основного тона, как показано на рис. 7.



а *б*

Рис. 7. Временное масштабирование:
а – входной речевой сигнал; *б* – речевой сигнал с измененным масштабом времени

Точная оценка параметров модели требует разделения сигнала на отдельные гармоники. Для этого после процедуры временного масштабирования, в результате которого частота основного тона становится постоянной, используется ДПФ-модулированный банк фильтров с числом каналов N_{f0}. В соответствии с теоремой Котельникова максимальное число анализируемых гармоник K определяется числом отсчетов на один период $K = N_{f0}/2$. Центры полос равнополосного банка фильтров, включающего N_{f0} каналов, являются кратными постоянной частоте основного тона. Коэффициенты усиления, характеризующие спектральную огибающую сигнала, определяются как значения мгновенных амплитуд субполосных сигналов $s_1(q),...,s_K(q)$.

После обработки параметров речевого сигнала выполняется синтез, который состоит из следующих шагов: 1) для каждой гармоники генерируется децимированная последовательность возбуждения; 2) для каждой гармоники вычисляется коэффициент усиления в соответствии с новыми параметрами модели; 3) субканальные сигналы возбуждения умножаются на коэффициенты усиления и пропускаются через банк фильтров синтеза для подавления эффекта наложения спектра, обусловленного децимацией; 4) масштаб времени сигнала изменяется в соответствии с целевым контуром частоты основного тона.

Результаты экспериментов

Описанный выше подход параметрического нестационарного моделирования речевых сигналов использован для решения следующих практических задач: синтез речи по тексту [5], коррекция певческого голоса [12], конверсия голоса [10, 11], изменение просодических характеристик речи [10] и кодирование звуковой информации [12–19]. Ниже приводятся результаты экспериментального сравнения предложенной системы моделирования с основными аналогичными решениями.

Для оценки точности предложенного алгоритма оценки мгновенной частоты основного тона используется набор искусственных, синтетических сигналов с заранее известными параметрами. Скорость изменения частоты основного тона тестовых сигналов изменяется от 0 до 2 Гц/мс. Значения мгновенной частоты находятся в пределах от 100 до 350Гц. Частота дискретизации сигналов – 44,1 кГц. Сравнивается пять различных алгоритмов: известные ранее RAPT, YIN, SWIPE' и две версии предложенного алгоритма оценки основного тона – одна без уточнения частоты основного тона (IRAPT 1) и вторая с уточнением частоты основного тона путем временного масштабирования сигнала (IRAPT 2).

К чистому тональному сигналу добавляется белый шум различной интенсивности для того, чтобы оценить устойчивость алгоритма к аддитивным шумам. Интенсивность шума определяется соотношением гармоники/шум (*HNR*)

$$HNR = 10\lg \sigma_H^2 \Big/ \sigma_N^2 ,$$

где σ_H^2 - энергия гармонического сигнала и σ_N^2 - энергия шума. Диапазон *HNR* изменяется от 25 дБ до 5дБ. Нижняя граница в 5дБ обусловлена тем, что фреймы с большим содержанием шума часто классифицируются RAPT как невокализованные.

Результат работы алгоритмов сравнивается в терминах 1) процент грубых ошибок (gross pitch error - GPE) и 2) средний процент мелких ошибок (mean fine pitch error – MFPE).

Процент грубых ошибок вычисляется как

$$GPE(\%) = \frac{N_{GPE}}{N_V} \times 100,$$

где N_{GPE} – число фреймов с отклонением полученной оценки более чем на $\pm 20\,\%$ от настоящего значения основного тона, N_V – общее число вокализованных фреймов.

Средний процент мелких ошибок вычисляется для вокализованных фреймов без грубых ошибок

$$MFPE(\%) = \frac{1}{N_{FPE}} \sum_{n=1}^{N_{FPE}} \frac{\left| F_0^{true}(n) - F_0^{est}(n) \right|}{F_0^{true}(n)} \times 100,$$

где N_{FPE} – число вокализованных фреймов без грубых ошибок, $F_0^{true}(n)$ – действительные значения основного тона и $F_0^{est}(n)$ – оценочные значения основного тона.

Результаты тестирования алгоритмов с использованием синтетических сигналов приведены в табл. 1.

Приведенные результаты экспериментов показывают, что все алгоритмы имеют низкие показатели GPE и MFPE в случае неизменной частоты основного тона и преимущество IRAPT 1-2 становится заметным с увеличением частотных модуляций – рис. 8.

При наличии белого шума высокой интенсивности предлагаемый алгоритм сохраняет свое преимущество, однако при низких значениях *HNR* версия IRAPT 1 может быть предпочтительнее чем IRAPT 2.

Таблица 1. **Сравнение алгоритмов оценки основного тона с использованием синтетических сигналов**

Алгоритм	Тип оценки	Скорость изменения частоты основного тона Гц/мс				
		0	0.5	1	1.5	2
HNR 25dB						
RAPT	GPE	**0**	**0**	**0**	7,90	18,42
	MFPE	0,037	0,103	0,219	0,405	0,778
YIN	GPE	**0**	**0**	**0**	**0**	5,36
	MFPE	**0,002**	0,156	0,778	2,136	3,905
SWIPE’	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,09	0,150	0,337	0,607	1,206
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,111	0,094	0,100	0,104	0,255
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,013	**0,050**	**0,051**	**0,060**	**0,114**
HNR 15dB						
RAPT	GPE	**0**	**0**	**0**	7,90	18,42
	MFPE	0,053	0,108	0,217	0,415	0,778
YIN	GPE	**0**	**0**	**0**	**0**	5,16
	MFPE	**0,004**	0,154	0,785	2,103	3,803
SWIPE’	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,165	0,193	0,347	0,632	1,194
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,113	0,094	0,102	0,111	0,273
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,049	**0,056**	**0,65**	**0,074**	**0,148**
HNR 5dB						
RAPT	GPE	**0**	**0**	**0**	10,52	18,42
	MFPE	0,161	0,205	0,268	0,506	0,871
YIN	GPE	**0**	**0**	**0**	**0**	4,33
	MFPE	**0,019**	0,151	0,813	1,948	3,524
SWIPE’	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,316	0,253	0,373	0,706	1,307
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,143	**0,099**	**0,115**	**0,147**	0,356
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0,162	0,131	0.145	0,164	**0,256**

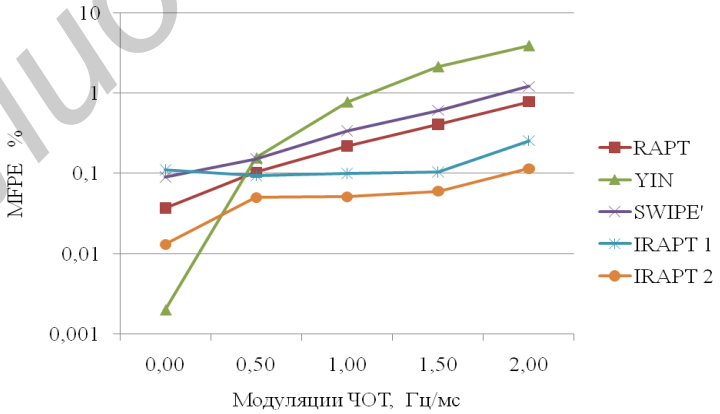


Рис. 8. Изменение точности оценки основного тона с увеличением частотных модуляций

Работа алгоритмов сравнивается с использованием натуральной речи при помощи речевой базы данных PTDB-TUG. База данных содержит 2342 предложения, взятых из речевого корпуса TIMIT, прочитанных 10 дикторами мужчинами и 10 дикторами женщинами. База данных включает контрольные сигналы, полученные при помощи ларингографа и их оценочные значения частоты основного тона. Данные значения не могут рассматриваться как мгновенные, поэтому нельзя сравнить алгоритмы так же достоверно как в случае с

синтетическими сигналами, однако эксперимент позволяет оценить применимость предложенного алгоритма к обработке настоящих речевых сигналов. Полученные результаты приведены в табл. 2.

Таблица 2. **Сравнение алгоритмов оценки частоты основного тона с использованием речевых сигналов**

Алгоритм	Мужской голос		Женский голос	
	GPE	MFPE	GPE	MFPE
RAPT	3,687	1,737	6,068	1,184
YIN	3,184	**1,389**	3,960	0,835
SWIPE’	**0,783**	1,507	4,273	**0,800**
IRAPT 1	1,625	1,608	**3,777**	0,977
IRAPT 2	1,571	1,565	**3,777**	1,054

Для натуральных речевых сигналов предложенный алгоритм показывает близкий результат к другим алгоритмам оценки, что говорит о его применимости в реальных приложениях обработки речи.

Предложенная модель речевого сигнала сравнивается с известной гибридной моделью речевого сигнала TANDEM-STRAIGHT при помощи средних значений экспертных оценок MOS (Mean Opinion Score). Для сравнения использовались речевые записи из базы данных CMU ARCTIC: два мужских голоса ('bdl' и 'rms') и два женских ('clb' и 'slt'). Выполняется параметрическое моделирование речи с использованием предложенной модели (обозначенной как 'GUSLY') и TANDEM-STRAIGHT (обозначенной как 'T-S'). Некоторые результаты моделирования доступны в интернете по адресу http://dsp.tut.su/gusly_vs_straight.rar.

В прослушивании участвовало несколько специалистов, которые оценивали качество обработанной речи по пятибалльной шкале (5: отлично, 4: хорошо, 3: нормально, 2: недостаточно, 1: плохо).

В первом эксперименте выполняется растяжение речевого сигнала по времени в 1,5 и 2,2 раза (коэффициенты растяжения обозначены 'x 1.5' и 'x 2.2' соответственно) с сохранением исходного тона. Результат эксперимента показан на рис. 9 (мужские голоса обозначены 'm', а женские 'f'). Видно, что предложенный метод превосходит TANDEM-STRAIGHT для коэффициента растяжения 1.5. Однако, при растяжении в 2.2 раза GUSLY показывает не такой высокий результат, что объясняется появлением эффекта 'опережающего эхо' транзиентных звуков речи.

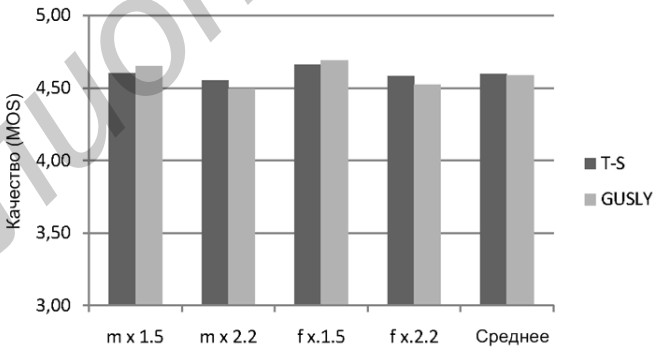


Рис. 9. Растяжение речевого сигнала по времени. Оценки MOS

Во втором эксперименте выполняется повышение основного тона сигнала с сохранением длительности воспроизведения и темпа произношения. Значения исходного основного тона умножаются на коэффициенты 1.2 и 1.9 (обозначенные '↑ 1.2' и '↑ 1.9' соответственно). Результаты эксперимента приведены на рис. 10. Для всех голосов результаты полученные при помощи модели GUSLY, превосходят результаты, полученные при помощи модели TANDEM-STRAIGHT.

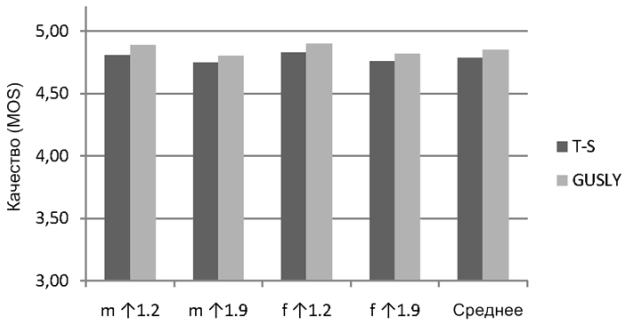


Рис. 10. Повышение основного тона. Оценки MOS

В третьем эксперименте выполняется понижение основного тона в 1/1.2 и 1/1.9 раза (коэффициенты понижения тона обозначены '↓ 1/1.2' и '↓ 1/1.9' соответственно). По результатам прослушивания, приведенным на рис. 11 видно, что модель GUSLY имеет оценки немного ниже, чем модель TANDEM-STRAIGHT. Это объясняется тем, что при понижении основного тона число гармоник, помещающихся в частотный диапазон сигнала, увеличивается и предложенная модель не имеет возможности оценить сигналы возбуждения для появившихся высокочастотных гармоник корректно.

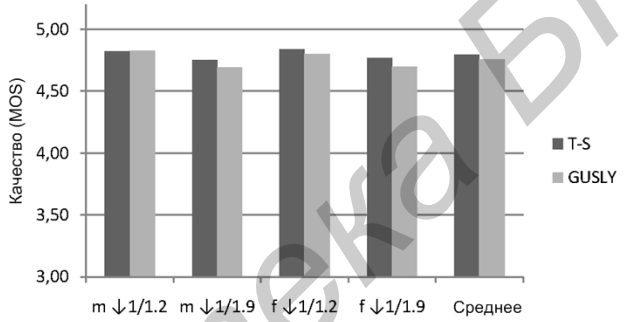


Рис. 11. Понижение основного тона. Оценки MOS

Заключение

В работе приведено краткое описание методов нестационарной параметризации речевых сигналов, позволяющих выполнять сложную обработку. Основным направлением дальнейшего исследования является поиск высокоуровневой модели, обеспечивающей адекватное описание процесса речеобразования с учетом индивидуальных особенностей речевого тракта, голосовых связок и артикуляции. Модель может существенно усовершенствовать полученные прикладные решения и будет полезной в задачах глубокой компрессии речи и верификации диктора.

TIME-VARYING PARAMETRIC REPRESENTATION OF SPEECH FOR MULTIMEDIA APPLICATIONS

A.A. PETROVSKY, I.S. AZAROV

Abstract

Methods of time-varying speech parameterization for analysis, processing and synthesis in multimedia systems. The main theoretical points are given and practical issues are discussed. Some practical results of instantaneous pitch estimation and quality of voice morphing are presented.

Список литературы

1. *Азаров И.С., Петровский А.А.* // Информатика. 2008. № 2. С. 71–82.
2. *Азаров И.С., Петровский А.А.* // Цифровая обработка сигналов. 2012. № 2. С. 15–23.
3. *Азаров И.С., Петровский А.А.* // Мгновенный гармонический анализ: обработка звуковых и речевых сигналов в системах мультимедиа. Саарбрюкен, 2011
4. *Azarov E., Petrovsky A.* // Recent advances in signal processing. Vienna, 2009.
5. *Petrovsky A., Azarov E., Petrovsky A.A.* // Signal processing. Munich, 2009.
6. *Азаров И.С., Петровский А.А.* // Речевые технологии. 2008. № 1 (1). С. 67–77.
7. *Азаров И.С., Петровский А.А.* // Докл. БГУИР. 2008. № 4 (34). С. 92–105.
8. *Петровский А.А, Азаров И.С.* Анализаторы речевых и звуковых сигналов: методы, алгоритмы и практика (с MATLAB примерами). Минск, 2009.
9. *Азаров И.С., Вашкевич М.И., Петровский А.А.* // Цифровая обработка сигналов. 2012. № 4. С. 49–57.
10. *Azarov E., Petrovsky A., Zubrycki P.* // Elektronika, PAN. 2011. № 5. P. 111–116.
11. *Azarov E., Petrovsky A., Parfieniuk M.* // EURASIP Journal on Advances in Signal Processing. 2010. Article ID 712749. P. 1–10.
12. *Petrovsky Al., Azarov E., Petrovsky A.* // Signal Processing. 2011. Vol. 91. Iss. 6. P. 1489–1504.
13. *Piotrowsk A., Parfieniuk M.* // Cyfrowe banki filtrow: analiza, synteza I implementacja dla systemow multimedialnych. Bialystok, 2006. .
14. *Zubrycki P., Pavlovec A., Petrovsky A.* New trends in audio and video. Vol. 1. Bialystok, 2006. P. 233–246.
15. *Вашкевич М.И., Петровский А.А.* // Докл. БГУИР. 2009. № 4. С. 5–10.
16. *Павловец А.Н., Лившиц М.З., Лихачев Д.С., Петровский А.А.* // Речевые технологии. 2008. № 4. С. 37–49.
17. *Павловец А.Н., Петровский А.А.* // Речевые технологии. 2008. № 4. С. 50–60.
18. *Лихачев Д.С., Азаров И.С., Петровский А.А.* // Информатика. 2011. № 4. С. 59–70.
19. *Parfieniuk M., Petrovsky A.A.* // INTL journal of electronics and telecommunications. 2012. Vol. 58. № 2. P. 177–192.

СВЕДЕНИЯ ОБ АВТОРАХ

Петровский Александр Александрович (1953 г.р.), д.т.н., профессор. В 1975 г. закончил с отличием МРТИ. В 1980 г. защитил кандидатскую диссертацию в МРТИ, в 1989 г. – докторскую диссертацию в Институте проблем моделирования в энергетике АН Украины. С мая 1990 г. занимает должность заведующего кафедрой ЭВС (ранее –КиП ЭВА). Научный руководитель НИЛ 3.1 «Мультипроцессорные системы реального времени». Главные научные интересы – цифровая обработка сигналов речи и звука для целей компрессии, распознавания, редактирования шума в сигнале, синтеза цифровых банков фильтров.

Азаров Илья Сергеевич (1980 г.р.), к.т.н., доцент. В 2002 г. окончил БГУ. В 2009 г. защитил кандидатскую диссертацию в БГУИР. С 2009 г. занимает должность доцента кафедры электронных вычислительных средств. В 2011 г поступил в докторантуру БГУИР. Область научных интересов – цифровая обработка сигналов, кодирование речи, синтез речи по тексту, конверсия голоса. Им опубликовано 12 статей в отечественных и зарубежных научных журналах и 1 монография.