УДК 004.934

Сухостат Л.В.
Институт Информационных Технологий НАНА, Баку, Азербайджан
lsuhostat@hotmail.com

ОБ ОДНОМ ПОДХОДЕ ДЛЯ НАХОЖДЕНИЯ ПЕРИОДА ОСНОВНОГО ТОНА РЕЧЕВЫХ СИГНАЛОВ С ПРИМЕНЕНИЕМ АДАПТИВНЫХ ВЕЙВЛЕТОВ

Среди существующих методов, применяемых при распознавании личности по голосу, лишь некоторые могут работать для нелинейных и нестационарных речевых сигналов. Период основного тона является одним из важных признаков, характеризующих говорящего. В данной работе представлен метод нахождения периода основного тона для нелинейных и нестационарных речевых сигналов на основе эмпирического вейвлет-преобразования. Эксперименты показывают достаточно высокую эффективность предлагаемого подхода при различных уровнях шума.

Ключевые слова: *период основного тона, эмпирическое вейвлет-преобразование, оператор разделения энергии Тигера-Кайзера, внутренняя модовая функция, мгновенная частота.*

Введение

Начиная с 1970-х годов было разработано множество алгоритмов для нахождения периода основного тона. Среди них можно выделить модифицированный автокорреляционный метод (Modified Autocorrelation Method, AUTOC) [1], кепстральный метод, многополосную агрегацию коррелограмм (Multi-Band Summary Correlogram, MBSC) [2], BaNa [3], YIN [4], YAAPT [5], среднее значение разностной функции (Average Magnitude Difference Function, AMDF) [1], SWIPE' [6] и метод оценки основного тона на основе амплитудного сжатия (Pitch Estimation Filter with Amplitude Compression, PEFAC) [7].

Методы нахождения периода основного тона можно, в общем, разделить на временные, частотные и гибридные. Первая категория, в основном, ищет пики в автокорреляционной функции, вторая – в спектре мощности, тогда как третья группа выполняет частотно-временной анализ выхода банка полосовых фильтров. Во многих случаях алгоритмы выделяют кандидатов периода основного тона для каждого временного фрейма, а затем используют временные ограничения на непрерывность.

Однако только некоторые из них могут работать для случая нелинейных и нестационарных сигналов. Основная причина состоит в том, что методы нахождения периода основного тона основаны на предположении, что процесс речеобразования линеен, а речевые сигналы являются локально стационарными.

Выбор алгоритма для нахождения периода основного тона всегда состоит в нахождении компромисса между временем и разрешающей способностью по частоте (frequency resolution), робастностью, задержкой и вычислительной сложностью.

Некоторые хорошие альтернативы – методы на основе оценки параметров мгновенной частоты. Мгновенная частота была представлена в [8, 9], оценка методов предложена в [8–13]. Она необходима для понимания подробных механизмов обработки нелинейных и нестационарных процессов. На практике мгновенная частота обычно вычисляется из внутренней модовой функции (Intrinsic Mode Function, IMF) с помощью преобразования Гильберта-Хунга (Hilbert-Huang Transform, HHT). HHT для получения IMF применяет эмпирическую модовую декомпозицию (Empirical Mode Decomposition, EMD) [13]. В сравнении с методом автокорреляции данный подход более точен и быстрее находит период основного тона. Алгоритм EMD, применяемый при вычислении IMF, адаптивен, его вычислительная сложность, включая число извлеченных IMF и количество

вычислений на этапе их получения, строго зависит от сложности самих речевых сигналов. Но метод EMD автоматически оценивает число мод, что существенно влияет на вычислительную сложность. В связи с этим в работе предлагается метод эмпирического вейвлет-преобразования (Empirical Wavelet Transform, EWT) [14].

Далее для получения мгновенной частоты применяется преобразование Гильберта. Однако оператор разделения энергии Тигера-Кайзера (Teager-Kaiser Energy Operator, TKEO) [15] превосходит его по вычислительной сложности и быстроте на реальных сигналах. Преобразование Гильберта и оператор TKEO могут быть применены только к монокомпонентным сигналам. В случае мультикомпонентных сигналов необходимо разбиение сигнала на простые компоненты перед применением методов. Для этого удобно использовать узкополосные фильтры [10]. Однако в случае частотно-модулированных компонентов это не всегда возможно в силу широкого диапазона частот.

В данной работе представлен новый метод нахождения периода основного тона на основе мгновенной частоты с применением EWT и нелинейного оператора TKEO. Приводятся результаты практических экспериментов на речевой базе данных Keele [16]. Оценивается робастность предложенного метода к шуму.

Эмпирическое вейвлет-преобразование

В работе предлагается метод построения семейства вейвлетов, адаптированных к обрабатываемым сигналам. Рассматриваются реальные сигналы, где спектр симметричен относительно частоты $\omega = 0$, а также нормализованная ось Фурье, которая имеет периодичность для того, чтобы удовлетворить критерию Шеннона, и ограничивается отрезком $\omega \in [0, \pi]$.

Отрезок $[0, \pi]$ делится на N смежных сегментов. Обозначим через ω_n границы между сегментами (где $\omega_0 = 0$ и $\omega_N = \pi$). Каждый сегмент $\Lambda_n = [\omega_{n-1}, \omega_n]$. Вокруг ω_n определяется переходная фаза T_n шириной $2\tau_n$.

Эмпирические вейвлеты [14] определяются как полосовые фильтры на каждом Λ_n. Эмпирическая масштабируемая функция и эмпирические вейвлеты определяются следующим образом:

$$\hat{\phi}_n(\omega) = \begin{cases} 1, & \text{если} \quad |\omega| \le \omega_n - \tau_n \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}\left(|\omega| - \omega_n + \tau_n\right)\right)\right], & \text{если} \quad \omega_n - \tau_n \le |\omega| \le \omega_n + \tau_n \\ 0, & \text{в противном случае} \end{cases} \qquad (1)$$

и

$$\hat{\psi}_n(\omega) = \begin{cases} 1, & \text{если} \quad \omega_n + \tau_n \le |\omega| \le \omega_{n+1} - \tau_{n+1} \\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_{n+1}}\left(|\omega| - \omega_{n+1} + \tau_{n+1}\right)\right)\right], & \text{если} \quad \omega_{n+1} - \tau_{n+1} \le |\omega| \le \omega_{n+1} + \tau_{n+1} \\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}\left(|\omega| - \omega_n + \tau_n\right)\right)\right], & \text{если} \quad \omega_n - \tau_n \le |\omega| \le \omega_n + \tau_n \\ 0, & \text{в противном случае} \end{cases} . \qquad (2)$$

Функция $\beta(x)$, произвольная из $C^k([0,1])$, такая, что

$$\beta(x)=\begin{cases}0, & если \quad x\le 0\\ 1, & если \quad x\ge 1\end{cases} \quad и\ \beta(x)+\beta(1-x)=1 \quad \forall x\in[0,1]. \tag{3}$$

τ_n выбирается пропорционально ω_n: $\tau_n=\gamma\omega_n$, где $0<\gamma<1$. Следовательно, для всех $n>0$ уравнения (1) и (2) принимают вид

$$\hat{\phi}_n(\omega)=\begin{cases}1, & если \quad |\omega|\le(1-\gamma)\omega_n\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & если \quad (1-\gamma)\omega_n\le|\omega|\le(1+\gamma)\omega_n\\ 0, & в\ противном\ случае\end{cases} \tag{4}$$

и

$$\hat{\psi}_n(\omega)=\begin{cases}1, & если \quad (1+\gamma)\omega_n\le|\omega|\le(1-\gamma)\omega_{n+1}\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_{n+1}}\left(|\omega|-(1-\gamma)\omega_{n+1}\right)\right)\right], & если \quad (1-\gamma)\omega_{n+1}\le|\omega|\le(1+\gamma)\omega_{n+1}\\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\gamma\omega_n}\left(|\omega|-(1-\gamma)\omega_n\right)\right)\right], & если \quad (1-\gamma)\omega_n\le|\omega|\le(1+\gamma)\omega_n\\ 0, & в\ противном\ случае\end{cases} . \tag{5}$$

Теперь можем определить эмпирическое вейвлет-преобразование $W_f^{\varepsilon}(n,t)$, так же как и в случае классического вейвлет-преобразования:

$$W_f^{\varepsilon}(n,t)=\langle f,\psi_n\rangle=\int f(\tau)\overline{\psi_n(\tau-t)}d\tau=\left(\hat{f}(\omega)\overline{\hat{\psi}_n(\omega)}\right)^{\vee}, \tag{6}$$

а аппроксимирующие коэффициенты – скалярных произведений с масштабируемой функцией

$$W_f^{\varepsilon}(0,t)=\langle f,\phi_1\rangle=\int f(\tau)\overline{\phi_1(\tau-t)}d\tau=\left(\hat{f}(\omega)\overline{\hat{\phi}_1(\omega)}\right)^{\vee}, \tag{7}$$

где $\hat{\psi}_n(\omega)$ и $\hat{\phi}_1(\omega)$ определяются из уравнений (4) и (5) соответственно. Обратное преобразование принимает вид

$$f(t)=W_f^{\varepsilon}(0,t)*\phi_1(t)+\sum_{n=1}^{N}W_f^{\varepsilon}(n,t)*\psi_n(t)=\left(\hat{W}_f^{\varepsilon}(0,\omega)*\hat{\phi}_1(\omega)+\sum_{n=1}^{N}\hat{W}_f^{\varepsilon}(n,\omega)*\hat{\psi}_n(\omega)\right)^{\vee}. \tag{8}$$

Функция IMF f_k определяется следующим образом:

$$f_0(t)=W_f^{\varepsilon}(0,t)*\phi_1(t), \tag{9}$$

$$f_k(t)=W_f^{\varepsilon}(k,t)*\psi_k(t). \tag{10}$$

Преобразование Гильберта

После получения IMF-компонент для вычисления мгновенной частоты и мгновенной амплитуды к каждой IMF применяется преобразование Гильберта [17]. Использование преобразования позволяет получить для каждого момента времени

$$H[c_j(t)] = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{c_j(\tau)}{t-\tau} d\tau \,. \tag{11}$$

Теперь можно построить аналитический сигнал $z_j(t)$ как

$$z_j(t) = c_j(t) + i\, H[c_j(t)], \tag{12}$$

который может быть представлен в виде

$$z_j(t) = \alpha_j(t) \exp(i\omega_j(t)) \,. \tag{13}$$

Мгновенная амплитуда $\alpha_j(t)$ и фазовая функция $\theta_j(t)$ определяются как

$$\alpha_j(t) = \sqrt{c_j^2(t) + H^2[c_j(t)]} \,, \tag{14}$$

$$\theta_j(t) = \arctan \frac{H[c_j(t)]}{c_j(t)} \,. \tag{15}$$

Кроме того, мгновенная частота $\omega_j(t)$ может быть вычислена следующим образом:

$$\omega_j(t) = \frac{d\theta_j(t)}{dt} \,. \tag{16}$$

Так, исходный сигнал может быть представлен в форме (5):

$$X(t) = \mathrm{Re} \sum_{j=1}^{n} \alpha_j(t) \exp[i \int \omega_j(t) dt], \tag{17}$$

где остаток опущен, а $\mathrm{Re}\{\cdot\}$ обозначает реальную часть комплексного выражения.

Выражение (17) позволяет представить мгновенную амплитуду и частоту в трехмерном пространстве, где амплитуда – это высота в частотно-временной плоскости. Это частотно-временное распределение представлено как спектр Гильберта $H(\omega,t)$:

$$H(\omega,t) = \mathrm{Re} \sum_{j=1}^{n} \alpha_j(t) \exp[i \int \omega_j(t) dt] \,. \tag{18}$$

ННТ удовлетворяет требованию адаптивности для анализа нестационарных сигналов. Таким образом, сигнал может быть локально и точно отображен во временной частотной области путем применения спектра Гильберта.

Оператор разделения энергии Тигера-Кайзера

Всем речевым файлам были присвоены имена с уникальным идентификационным кодом, которые содержат поля, соответствующие идентификатору диктора, сессии и полу диктора. Оператор разделения энергии Тигера-Кайзера – нелинейный оператор, который успешно применяется во многих инженерных приложениях [15]. Он обнаруживает

модуляцию энергии и определяет мгновенную частоту и мгновенную амплитуду от AM-FM сигнала [9]. Оператор TKEO $\psi(\cdot)$ для сигнала $x(t)$ определяется как

$$\psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \tag{19}$$

где $\dot{x}(t)$ и $\ddot{x}(t)$ – производные первого и второго порядка соответственно. В дискретно-временной области оператор принимает вид

$$\psi[x(n)] = x^2(n) - x(n+1)\cdot x(n-1). \tag{20}$$

Мгновенная частота $f(n)$ и мгновенная амплитуда $|\alpha(n)|$ в любой момент времени для сигнала $x(n)$ даются как

$$y(n) = x(n) - x(n-1), \tag{21}$$

$$f(n) = \arccos\left(1 - \frac{\psi[y(n)] + \psi[y(n+1)]}{4\psi[x(n)]}\right), \tag{22}$$

$$|\alpha(n)| = \sqrt{\frac{\psi[x(n)]}{\sin^2[f(n)]}}, \tag{23}$$

$$f(n) = \frac{1}{2}\arccos\left(1 - \frac{\psi[x(n+1) - x(n-1)]}{2\psi[x(n)]}\right), \tag{24}$$

$$|\alpha(n)| = \frac{2\psi[x(n)]}{\sqrt{\psi[x(n+1) - x(n-1)]}}. \tag{25}$$

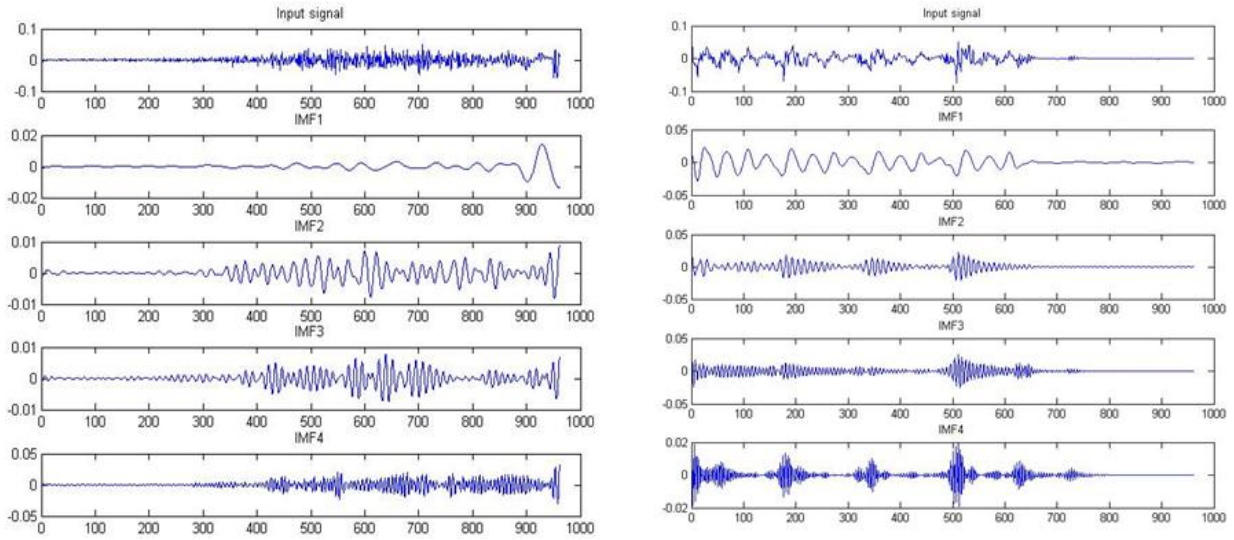
В общем, метод демодуляции (21–23) известен как первый дискретный алгоритм разделения энергии (DESA-1), а метод (24) и (25) – как второй дискретный алгоритм разделения энергии (DESA-2). Алгоритм DESA-2 требует только три значения для вычисления энергии в каждый момент времени и более прост в вычислении. Поэтому в данной работе мы рассматриваем алгоритм DESA-2.

Результаты экспериментов

Для проведения экспериментов была рассмотрена речевая база данных Keele [16]. Речевые образцы получены от 10 дикторов (5 мужчин и 5 женщин). Также содержатся записи, извлеченные из ларингографа (laryngograph) с помощью алгоритма автокорреляции и сопровождаются измерениями F_0.

Для тестирования устойчивости к внешним шумам рассматриваемых алгоритмов определения периода основного тона к сигналам добавляется белый шум при различных соотношениях сигнал-шум (Signal-to-Noise Ratio, SNR). Для генерации зашумленной речи с определенным значением SNR энергия сигнала вычисляется только на вокализованных участках речевого сигнала, и шум усиливается или ослабевает до определенного уровня, чтобы удовлетворить значению целевого SNR.

Период основного тона был оценен с помощью автокорреляции с окном 26,5 мс и сдвигом в 10 мс.



а) б)

Рис. 1. Декомпозиция сигналов с помощью EWT

Для сравнения производительности методов нахождения периода основного тона используются следующие метрики ошибок [17]:

а) Процент грубых ошибок (Gross Pitch Error, GPE) определяет соотношение кадров, рассчитанное на основе вокализованных участков полученного периода основного тона и эталонных значений (ground truth), для которых относительная погрешность оценки выше, чем определенный порог δ (обычно 20% [18]):

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\% \ , \tag{26}$$

где N_{VU} – число фреймов, в которых полученное значение основного тона и эталонное значение определены как вокализованные, N_{F0E} – число фреймов, для которых выполняется условие

$$\left| \frac{F0_{est}(i)}{F0_{true}(i)} - 1 \right| > \delta\% \, , \tag{27}$$

где $F0_{true}(i)$ – эталонное значение периода основного тона, а $F0_{est}(i)$ – полученное значение периода основного тона, i – число фреймов.

б) Средний процент мелких ошибок (Mean Fine Pitch Error, MFPE) вычисляется на вокализованных участках, где не наблюдаются ошибки GPE [19].

$$MFPE = \frac{1}{N_{FPE}} \sum_{i=1}^{N_{FPE}} \frac{\left| F0_{true}(i) - F0_{est}(i) \right|}{F0_{true}(i)} \times 100\% \, , \tag{28}$$

где N_{FPE} – число вокализованных участков без GPE.

EWT разлагает сегменты сигнала на серию IMF-функций для дальнейшего извлечения мгновенной частоты (рис. 1). Из всего набора IMF выбираем ту, чей период ближе к исходному сигналу. Из рис. 1 видно, что IMF1 содержит информацию о периоде основного тона и хорошо отображает форму сигнала. Блок-схема процесса на основе EWT-TKEO показана на рис. 2 (здесь IF1 обозначает мгновенную частоту, получаемую из первой IMF).

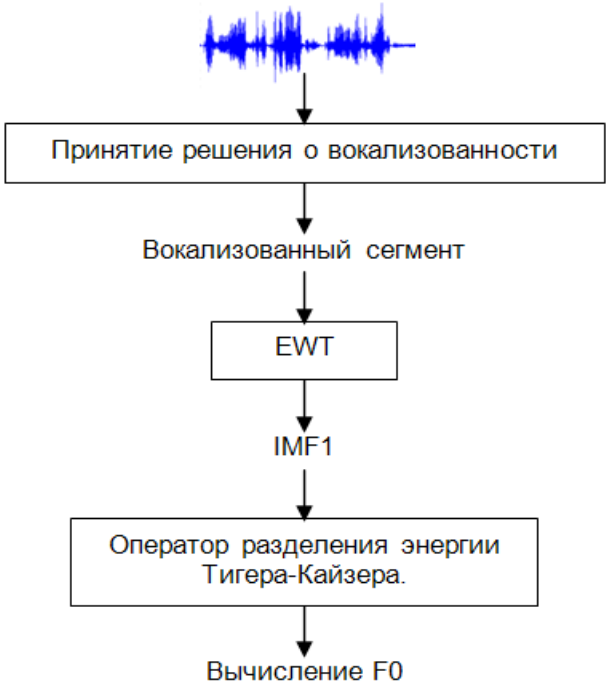


Рис. 2. Схема нахождения периода основного тона

Эксперименты были проведены в среде Matlab 2011b. Сравнение производительности предложенного метода и других популярных алгоритмов нахождения периода основного тона приводится в таблице 1.

Таблица 1

Сравнение производительности методов нахождения периода основного тона

SNR (dB) / Метод		-5	0	10	15
SWIPE’	GPE	20,82	14,28	11,92	8,43
	MFPE	1,77	1,21	0,90	0,83
YAAPT	GPE	13,90	23,35	9,98	6,26
	MFPE	2,10	1,94	1,51	0,81
HHT	GPE	19,30	6,02	11,34	5,10
	MFPE	0,56	0,43	0,35	0,33
EWT-TKEO	GPE	15,11	5,77	10,56	4,79
	MFPE	0,37	0,30	0,28	0,14

Как показано в таблице 1, GPE для предложенного метода на основе EWT по сравнению с другими методами извлечения периода основного тона значительно меньше, чем у методов EMD, YIN и SWIPE’ при различных уровнях шума. Предложенный метод является более робастным по сравнению с другими методами и лучше работает при высоких SNR.

Заключение

Целью текущего исследования была разработка метода нахождения периода основного тона речевого сигнала. Идея предлагаемого подхода состоит в том, что мгновенная частота содержит информацию о периоде основного тона. Был рассмотрен метод EWT. Для выделения мгновенной частоты был предложен оператор TKEO. Для проведения экспериментов белый шум был добавлен к речевому сигналу перед применением метода извлечения периода основного тона. Алгоритм протестирован при различных уровнях шума. Было показано, что точность алгоритма выше, чем у алгоритмов YAAPT, EMD и SWIPE' в случае быстрых модуляций высоты. Эксперименты показывают достаточную эффективность предлагаемого подхода при аддитивном шуме.

Литература

1. Rabiner L.A., Cheng M.J., Rosenberg A.E., McGonegal C.A. A comparative performance study of several pitch detection algorithms // IEEE Trans. on Acoust., Speech and Signal Proc., 1976, no.5, pp.399–417.
2. Tan L.N., Alwan A. Multi-band summary correlogram-based pitch detection for noisy speech // Speech Communication, 2013, vol.55, no.78, pp.841–856.
3. Ba H., Yang N. BaNa: a hybrid approach for noise resilient pitch detection // IEEE Statistical Signal Processing Workshop, 2012, pp.369–372.
4. De Cheveigne A., Kawahara H. Yin, a fundamental frequency estimator for speech and music // J. Acoust. Soc. Am., 2002, vol.111, no.4, pp.1917–1930.
5. Kasi K., Zahorian S.A. Yet another algorithm for pitch tracking / Proc. of the ICASSP, 2002, pp.361–364.
6. Camacho A. SWIPE: a sawtooth waveform inspired pitch estimator for speech and music. Ph.D. dissertation. Florida, 2007, 116 p.
7. Gonzalez S., Brookes M. A pitch estimation filter robust to high levels of noise (PEFAC) / Proc. of EUSIPCO, 2011, pp. 451–455.
8. Boashash B. Estimating and interpreting the instantaneous frequency of a signal // Proc. IEEE, 1992, vol.80, no.4., pp.520–568.
9. Maragos P., Kaiser J.F., Quatieri T.F. On amplitude and frequency demodulation using energy operators // IEEE Trans. on Signal Processing, 1993, vol.41, no.4, pp.1532–1550.
10. Abe T., Kobayashi T., Imai S. Harmonics tracking and pitch extraction based on instantaneous frequency / Proc. of ICASSP, 1995, vol.1, pp.756–759.
11. Abe T., Honda M. Sinusoidal model based on instantaneous frequency attractors // IEEE Trans. on Audio, Speech and Language Processing, 2006, vol.14, no.4, pp.1292–1300.
12. Azarov E., Petrovsky A., Parfieniuk M. Estimation of the instantaneous harmonic parameters of speech / Proc. of EUSIPCO, 2008, pp.1–5.
13. Huang N.E., Shen Z., Long S.R., Wu M.L., Shih H.H., Zheng Q., Yen N.C., Tung C.C., Liu H.H. The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis // Proc. Roy. Soc. London A, 1998, vol.545, pp.903–995.
14. Gilles J. Empirical Wavelet Transform // IEEE Transactions on Signal Processing, 2013, vol.61, no.16, pp.3999–4010.
15. Vakman D. On the analytic signal, the Teager–Kaiser energy algorithm, and other methods for defining amplitude and frequency // IEEE Trans. on Signal Process., 1996, vol.44, no.4, pp.791–797.
16. Chu W., Alwan A. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend / Proc. of ICASSP, 2009, pp.3969–3972.

17. Varga A., Steeneken H.J. Assessment for automatic speech recognition: II. Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems // Speech Communication, 1993, vol.12, no.3, pp.247–251.
18. Drugman T., Alwan A. Joint robust voicing detection and pitch estimation based on residual harmonics / Proc. of Interspeech, 2011, pp.1973–1976.
19. Azarov E., Vashkevich M., Petrovsky A. Instantaneous pitch estimation based on RAPT framework / Proc. of EUSIPCO, 2012, pp.2787–2791.

UOT 004.934

Suxostat Lyudmila V.
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
lsuhostat@hotmail.com

Adaptiv veyvletin tətbiqi ilə səs siqnallarının əsas ton periodunun tapılması üçün bir yanaşma haqqında

Səsə görə şəxsin tanınması üçün tətbiq edilən mövcud metodların arasında yalnız bəziləri qeyri-xətti və qeyri-stasionar siqnallarla işləyə bilərlər. Əsas ton periodu danışanı xarakterizə edən əhəmiyyətli əlamətlərdən biridir. Bu işdə empirik veyvlet çevirməsi əsasında qeyri-xətti və qeyri-stasionar nitq siqnalları üçün əsas ton periodunun tapılması metodu təklif edilmişdir. Aparılmış eksperimentlər müxtəlif küy səviyyələrində təklif edilən yanaşmanın kifayət qədər yüksək effektivliyini göstərir.

Açar sözlər: *əsas ton periodu, empirik veyvlet çevirmə, Tiqer-Kayzerin enerjisinin bölməsinin operatoru, daxili mod funksiya, ani tezlik.*

Lyudmila V. Sukhostat
Institute of Information Technology of ANAS, Baku, Azerbaijan
lsuhostat@hotmail.com

An approach to pitch period detection of speech signal based on adapted wavelets

Among the existing methods used for speaker recognition, only a few can work in the case of non-linear and non-stationary speech signals. Pitch period is one of the most important features for speaker characterization. This paper presents a method for pitch period detection of nonlinear and non-stationary speech signals based on empirical wavelet transform. Experiments show high relative efficiency of the proposed approach for different noise levels.

Keywords: *pitch period, empirical wavelet transform, Teager-Kaiser energy operator, intrinsic mode function, instantaneous frequency.*