

Алгоритм измерения частоты основного тона речевых сигналов на основе комплементарной множественной декомпозиции на эмпирические моды

А. К. АЛИМУРАДОВ

Пензенский государственный университет,
Пенза, Россия, e-mail: alansapfir@yandex.ru

Рассмотрена проблема повышения точности измерений частоты основного тона речевых сигналов. Представлены существующие алгоритмы определения этой частоты, разработан новый алгоритм на основе метода комплементарной множественной декомпозиции на эмпирические моды. Результаты исследований подтверждают устойчивость алгоритма к частотным модуляциям основного тона речевых сигналов.

Ключевые слова: обработка речевых сигналов, комплементарная множественная декомпозиция на эмпирические моды, частота основного тона.

The problem of measurement of speech signals pitch frequency is discussed. The existing pitch frequency measurement algorithms are presented, and the need to improve the measurement accuracy using adaptive processing techniques is proved. A new PF measurement algorithm based on the method of Complementary Ensemble Empirical Mode Decomposition is developed. The experimental studies of the developed algorithm are conducted, and the results prove the algorithm stability to the pitch frequency modulation and increase in PF measurement accuracy.

Key words: speech signal processing, Complementary Ensemble Empirical Mode Decomposition, pitch frequency.

Речевой сигнал является нестационарным акустическим сигналом сложной формы, амплитудные и частотные характеристики которого быстро изменяются во времени. Такие сигналы состоят из вокализованных и шумовых участков, образующихся соответственно в результате периодических и непериодических колебаний голосовых связок. Частота колебаний связок — важный информативный параметр речевых сигналов. Этот параметр называется частотой основного тона (ЧОТ) [1].

В настоящее время известно большое количество способов определения ЧОТ, которые можно классифицировать как алгоритмы измерений во временной, частотной и частотно-временной областях [2—7]. Широкое практическое применение получили алгоритмы на основе автокорреляционной функции и её модификаций (YIN) [2], устойчивого метода отслеживания основного тона (Robust Algorithm for Pitch Tracking, RAPT) [5] и основного тона пилообразной формы (Sawtooth Waveform Inspired Pitch Estimator, SWIPE) [7].

Популярность данных алгоритмов обусловлена функциональностью, низким процентом грубых ошибок и существованием в открытом доступе программных реализаций. Однако возможности таких алгоритмов ограничены низким частотным и временным разрешением методов обработки, неадаптивных к нестационарным речевым сигналам. Поэтому при измерениях ЧОТ актуально применять адаптивный метод обработки нестационарных сигналов, а именно — комплементарную множественную декомпозицию на эмпирические моды (КМДЭМ) [8].

Цель настоящей работы — разработка алгоритма для измерений ЧОТ с использованием адаптивного метода КМДЭМ. Статья является развитием ранее опубликованных

трудов, посвящённых применению комплементарной декомпозиции в задачах обработки речевых сигналов [9, 10].

Методы декомпозиции на эмпирические моды. Декомпозиция представляет адаптивную технологию разложения сигнала на внутренние функции, называемые эмпирическими модами (ЭМ) [11]. Особенность декомпозиции заключается в том, что базисные функции, используемые для разложения, извлекают непосредственно из исходного сигнала. При декомпозиции модель сигнала заранее не задана, ЭМ вычисляют в ходе процедуры разложения с учётом локальных особенностей сигнала, таких как экстремумы и нули, а также внутренней структуры каждого конкретного сигнала. Таким образом, ЭМ не имеют строгого аналитического описания, но должны удовлетворять условиям, гарантирующим определённую симметрию и узкополосность базисных функций [11]: общее количество экстремумов соответствует общему числу нулей с точностью до единицы; среднее значение верхней и нижней огибающих, интерполирующих соответственно локальные максимумы и минимумы, должно быть приближённо равно нулю.

Аналитическое выражение декомпозиции на ЭМ имеет следующий вид:

$$x(t) = \sum_{i=1}^I IMF_i(t) + r_1(t),$$

где $x(t)$ — исходный сигнал; i — номер ЭМ; I — количество ЭМ; $IMF_i(t)$ — конечное число извлекаемых ЭМ; r_1 — результирующий остаток.

Недостатком такой декомпозиции, ограничивающим её практическое применение, является смешивание ЭМ, т. е. существование в моде участков сигнала, несоизмеримых по

амплитудному и частотному масштабам. Для решения данной проблемы предложен метод множественной декомпозиции на ЭМ (МДЭМ), основанный на многократном добавлении к исходному сигналу белого шума малой амплитуды и вычислении среднего значения полученных мод как конечного результата [12]. Анализируемый сигнал является суммой сигнала и шума:

$$y_j(t) = x(t) + w_j(t);$$

$$IMF_j(t) = \sum_{j=1}^J IMF_{ji}(t)/J;$$

$$r_1(t) = \sum_{j=1}^J r_{ji}(t)/J,$$

где $w_j(t)$ — белый шум; $IMF_{ji}(t)$, $r_{ji}(t)$ — соответственно ЭМ и результирующий остаток, полученные при различных декомпозициях; $j = 1, 2, \dots, J$ — количество циклов декомпозиций (добавлений белого шума к сигналу).

В данном подходе использовано преимущество статистических характеристик белого шума для обнаружения слабых периодических или квазипериодических составляющих сигналов. Однако использование белого шума в МДЭМ порождает появление в сигнале нового шумового остатка, влияющего на реконструкцию исходного сигнала [8]. Для устранения этого недостатка был предложен метод с добавлением белого шума с прямыми и инверсными значениями амплитуды, т. е. метод КМДЭМ [8]:

$$\begin{bmatrix} y_j(t) \\ y_j(t)^* \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \times \begin{bmatrix} x(t) \\ w_j(t) \end{bmatrix},$$

где $y_j(t)^*$ — речевой сигнал, зашумлённый инверсным по знаку белым шумом.

В результате разложения сигнала методом КМДЭМ формируется множество мод, свободных от отмеченных недостатков декомпозиции на ЭМ и МДЭМ. Вычисление среднего значения полученных мод обеспечит полное удаление остаточного белого шума за счёт пары прямых и инверсных значений независимо от того, сколько сигналов шума использовалось. Для реализации математических аппаратов МДЭМ и КМДЭМ важны также амплитуда добавляемого белого шума и количество циклов декомпозиции.

Алгоритм измерения ЧОТ речевых сигналов на основе КМДЭМ. В [13—15] представлен обзор алгоритмов определения ЧОТ, использующих инструменты анализа нестационарных данных. В этих работах использованы методы классической и множественной декомпозиций. В условиях выраженного эффекта смешивания мод и остаточного шума точность определения ЧОТ зависит от корректного определения ЭМ, содержащих основной тон.

В результате проведённых исследований разработан алгоритм измерений ЧОТ, основанный на комплементарной декомпозиции (рис. 1) [16, 19]. Суть алгоритма заключается в адаптивном разложении речевого сигнала на ЭМ с помощью метода КМДЭМ. Из полученного набора ЭМ после анализа логарифма энергий определяют моду, содержащую основной тон. Далее, применяя функцию оператора Тигра (Teager Energy Operator, TEO), измеряют частоту. Рассмотрим подробнее основные этапы работы алгоритма.

Регистрация речевого сигнала $x(n)$ осуществляется со следующими параметрами: длительность записи — не более 10000 мс; частота дискретизации 8000 Гц, разрядность квантования 16 бит. При этом $0 \leq n \leq N$, где n — номер дискретного отсчёта времени, N — количество отсчётов в сигнале.

Комплементарная множественная декомпозиция. Результатом декомпозиции речевого сигнала будет определённое количество ЭМ $IMF_i(n)$ и результирующий остаток $r_1(n)$.

Определение энергетических характеристик ЭМ. Амплитудное распределение сигналов найденных ЭМ во временной области можно описать с помощью функции кратковременной энергии от времени. В разработанном алгоритме для сжатия амплитуды сигнала при большом динамическом диапазоне применяют логарифмирование энергии, максимально приближая действие алгоритма к работе слухового аппарата человека:

$$LE_i = \log_2 \sum_{n=1}^N [IMF_i(n)]^2,$$

где LE_i — логарифм энергии ЭМ.

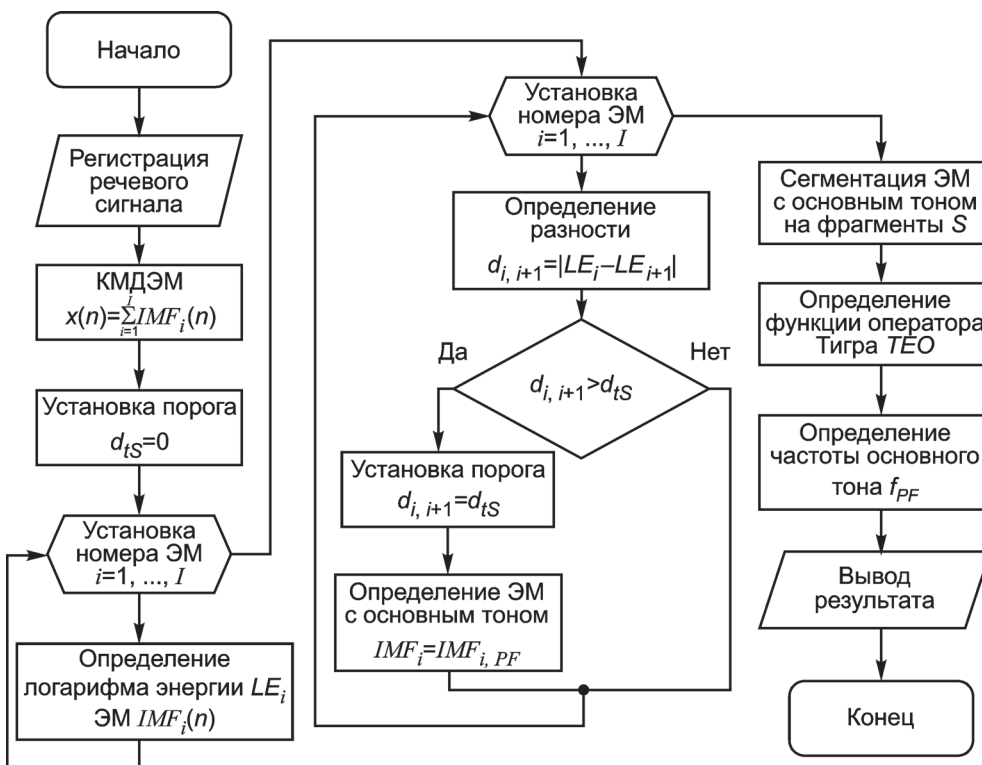


Рис. 1. Блок-схема алгоритма измерения частоты основного тона на основе комплементарной множественной декомпозиции на эмпирические моды

Определение ЭМ с основным тоном построено на предположении, что вокализованные моды имеют большую энергию, чем невокализованные. Процесс заключается в последовательном вычислении модуля разности значений логарифмов энергии текущей и последующей ЭМ:

$$d_{i,i+1} = |LE_i - LE_{i+1}|,$$

где $d_{i,i+1}$ — результат вычитания логарифмов энергии текущей и последующей ЭМ.

Используя пороговую обработку, определим из последовательности полученных данных наибольшее значение $d_{i,i+1}$. Этому значению соответствует резкий спад энергии между вокализованной ЭМ, содержащей основной тон, и невокализованной модой [14]. На рис. 2 представлена графическая интерпретация процесса определения моды, содержащей основной тон. В соответствии с упомянутым выше правилом определения из рис. 2, а следует, что пятая ЭМ содержит основной тон. Анализ спектрального распределения этой моды и исходного речевого сигнала подтверждает правильность определения моды, содержащей основной тон: единственная гармоническая составляющая ЧОТ пятой ЭМ соответствует первой составляющей формантных частот исходного речевого сигнала (рис. 2, б).

Сегментация ЭМ с основным тоном представляет линейное деление на составляющие отрезки — фрагменты. Разработанный алгоритм основан на предположении, что свойства речевого сигнала с течением времени изменяются медленно. Это допущение приводит к кратковременному анализу, в котором фрагменты ЭМ выделяют и обрабатывают так, как если бы они были короткими участками речевых сигналов с отличающимися свойствами. Сегментацию ЭМ на фрагменты осуществляют по следующим формулам:

$$S = IMF_i(n)/L;$$

$$IMF_{i,s+1}(n) = IMF_i[(s \cdot L) + 1; (s + 1) \cdot L],$$

где S — количество фрагментов в ЭМ; L — количество дискретных отсчётов в одном фрагменте; $IMF_{i,s+1}(n)$ — фрагмент i -й ЭМ; $s = (0, 1, 2, \dots, S-1)$ — номер фрагмента.

Определение ЧОТ осуществляют с использованием функции измерения мгновенной энергии сигнала, получившей название оператора Тигра и обладающей простотой, эффективностью и восприимчивостью к резкому изменению сигнала:

$$TEO = IMF_{i,PF}(n)^2 - IMF_{i,PF}(n-1) IMF_{i,PF}(n+1),$$

где TEO — функция оператора Тигра; $IMF_{i,PF}(n)$ — дискретное значение ЭМ, содержащей основной тон, в n -м отсчёте.

Для измерений ЧОТ используют близкорасположенные максимумы функции оператора Тигра (рис. 3):

$$f_{PF} = T_{\max, m+1} - T_{\max, m} / f_d,$$

где f_{PF} — ЧОТ; $T_{\max, m+1}$, $T_{\max, m}$ — максимумы функции оператора Тигра; m — номер максимума функции оператора Тигра; f_d — частота дискретизации.

Исследование алгоритма измерений ЧОТ речевых сигналов на основе КМДЭМ. Для оценки точности измерений ЧОТ в пакете прикладных

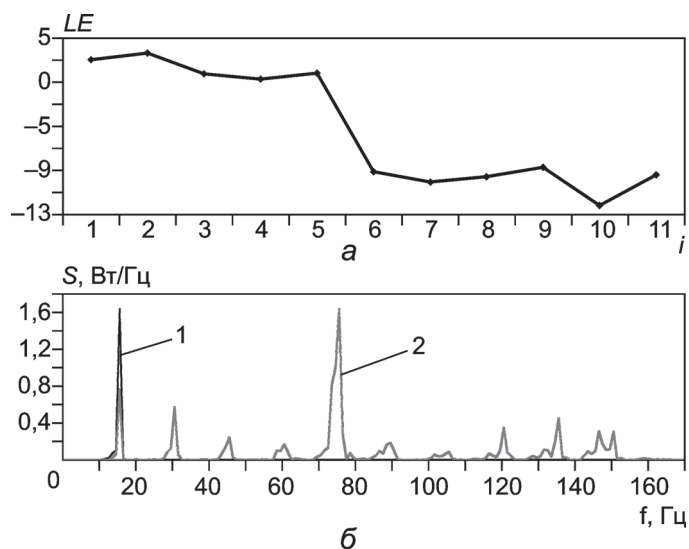


Рис. 2. Пример определения эмпирической моды, содержащей основной тон:

а — логарифмы энергии эмпирической моды; б — спектральное распределение пятой эмпирической моды (1) и исходного речевого сигнала (2)

программ MATLAB 7.0.1 провели экспериментальное исследование разработанного алгоритма. Методика эксперимента предполагала выполнение трёх этапов: формирование входного мультигармонического сигнала, измерение ЧОТ и сравнение результата с входным сигналом. В рамках исследования изменяли такие параметры входного мультигармонического сигнала, как частота сигнала основного тона f и скорость v изменения ЧОТ, а также параметр математического аппарата КМДЭМ — амплитуду A добавляемого белого шума. Количество циклов декомпозиции j было фиксировано и составляло 100.

В качестве критериев оценки использовали коэффициенты ошибок GPE и $MEPE$ (Gross Pitch Error и Mean Fine Pitch Error) [17]. Коэффициент GPE — безразмерная величина, равная отношению числа фрагментов S_{GRE} с отклонением измеренного значения ЧОТ более чем на 20 % от истинного значения этой частоты к общему числу фрагментов S_{PF} , содержащих основной тон:

$$GPE = (S_{GRE}/S_{PF}) \cdot 100.$$

Средний коэффициент $MFPE$ — тоже безразмерная величина, равная среднему значению отношения разности истинного f_{PFi} и измеренного значений f_{PF} ЧОТ к истинному значению фрагментов, содержащих основной тон без ошибок GPE :

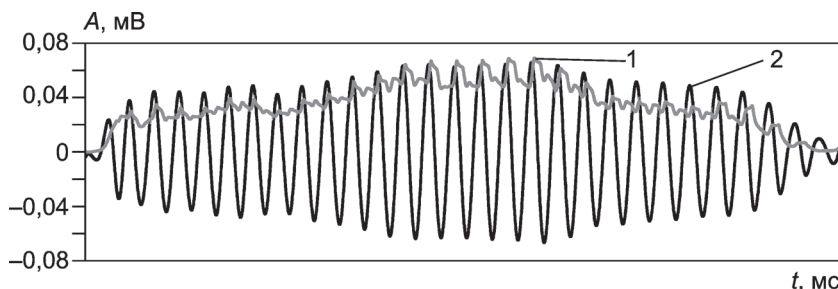


Рис. 3. Кривая функции оператора Тигра (1) и осциллограмма пятой эмпирической моды, содержащей основной тон (2)

$$MFPE = \frac{1}{S_{FPE}} \sum_{s=1}^{S_{FPE}} \left| \frac{f_{PFH}(s) - f_{PF}(s)}{f_{PFH}(s)} \right| \cdot 100,$$

где S_{FPE} — количество фрагментов, содержащих основной тон без ошибок GPE ; s — номер фрагмента, содержащего основной тон.

Для исследования сформировали тестовую выборку из 100 синтезированных мультigarмонических сигналов, представляющих сумму нескольких гармонических составляющих основного тона с заранее известной частотой. В соответствии с методикой исследований в рамках первого этапа каждый тестовый сигнал подвергался частотному модулированию — изменению ЧОТ в диапазоне 0—2,5 Гц/мс с шагом 0,5 Гц/мс. В рамках второго этапа методики в математическом аппарате КМДЭМ изменению подвергали амплитуду добавляемого белого шума с целью определения оптимального значения, при котором будут достигнуты наиболее точные результаты измерения ЧОТ.

В табл. 1 представлены результаты исследований разработанного алгоритма — зависимости коэффициента ошибок GPE и среднего коэффициента ошибок $MFPE$ от амплитуды добавляемого белого шума при различных скоростях v изменения ЧОТ.

Зависимости GPE и $MFPE$ от амплитуды добавляемого белого шума при различных скоростях v

v , Гц/мс	Критерии оценки, %	Амплитуда добавляемого белого шума A , мВ					
		0,10	0,25	0,50	1,00	2,00	3,00
0	GPE	0	0	0	0	0	0
	$MFPE$	1,68	1,20	1,31	1,37	2,67	3,95
0,5	GPE	0	0	0	0	0	0
	$MFPE$	4,61	4,50	4,61	5,27	5,86	6,21
1,0	GPE	0	0	0	0	0	0
	$MFPE$	5,37	5,23	5,31	5,87	6,31	6,78
1,5	GPE	0	0	0	0	0	0
	$MFPE$	6,41	6,23	6,39	6,92	7,24	7,85
2,0	GPE	0	0	0	0	0	5,4
	$MFPE$	8,94	8,75	8,83	9,21	9,84	10,26
2,5	GPE	0	0	0	0	7,90	8,60
	$MFPE$	10,35	10,14	10,24	10,90	11,25	11,60

Из табл. 1 следует, что самый точный результат измерений ЧОТ (выделено серым цветом) достигается при амплитуде $A = 0,25$ мВ. В соответствии с оптимальными параметрами математического аппарата КМДЭМ ($A = 0,25$ мВ, $j = 100$) провели исследование разработанного алгоритма на реальных речевых сигналах. Исходные данные: 100 речевых сигналов (5 тональных звуков «а», «о», «у», «е», «и» по 20 записей каждый), произнесённые 20 дикторами.

Результаты оценивали и сравнивали с рассмотренными ранее алгоритмами измерений ЧОТ: YIN , $RAPT$ и $SWIPE$ (табл. 2).

Таблица 2

Сравнительный анализ результатов измерения ЧОТ

v , Гц/мс	Критерии оценки	Алгоритмы измерений ЧОТ			
		YIN	$RAPT$	$SWIPE$	Разработанный алгоритм
0	GPE	0	0	0	0
	$MFPE$	3,10	2,56	2,35	1,10
0,5	GPE	0	0	0	0
	$MFPE$	6,56	6,65	6,34	3,21
1,0	GPE	0	0	0	0
	$MFPE$	7,30	7,25	6,81	6,54
1,5	GPE	0	0	0	0
	$MFPE$	8,21	8,50	7,90	8,00
2,0	GPE	2,05	0	0	0
	$MFPE$	10,20	10,10	9,67	9,54
2,5	GPE	7,60	5,30	6,20	6,50
	$MFPE$	14,20	13,10	13,50	13,20

Представленные результаты исследований показывают (выделено серым цветом), что при небольшой частотной модуляции разработанный алгоритм значительно превосходит существующие алгоритмы измерений ЧОТ. Однако при

Таблица 1

больших скоростях изменений частоты все алгоритмы имеют примерно соизмеримые значения GPE и $MFPE$. Это свидетельствует о том, что разработанный алгоритм может быть рекомендован для применения в задачах измерения ЧОТ.

Заключение. В статье рассмотрен вопрос повышения точности измерений ЧОТ речевых сигналов. Рассмотрены получившие широкое практическое применение алгоритмы измерений ЧОТ: YIN , $RAPT$ и $SWIPE$. Представлены результаты применения этих алгоритмов. Выявлена необходимость повышения точности измерений за счёт применения адаптивных методов обработки. Разработан алгоритм определения ЧОТ на основе адаптивного метода обработки нестационарных сигналов — КМДЭМ. Основным достоинством алгоритма является устойчивость к частотным модуляциям основного тона. Представлены блок-схема алгоритма и подробное математическое описание основных этапов работы.

Для оценки точности измерений ЧОТ проведено исследование разработанного алгоритма. В соответствии с полученными результатами сделан вывод, что алгоритм превосходит существующие аналоги при небольшой частотной модуляции.

Литература

1. Михайлов В. Г., Златоусова Л. В. Измерение параметров речи / Под ред. М. А. Сапожникова. М.: Радио и связь, 1987.
2. Camacho A., Harris J. G. A sawtooth waveform inspired pitch estimator for speech and music // J. Acoust. Soc. Am. 2008. V. 123 (4). N. 9. P. 1638—1652.

3. **Жиляков Е. Г., Фирсова А. А.** Оценивание периода основного тона звуков русской речи // Научные ведомости Белгородского государственного университета. Серия: История. Политология. Экономика. Информатика. 2013. № 1 (144). Вып. 25/1. С. 173—181.
4. **Azarov E., Vashkevich M., Petrovsky A.** Instantaneous pitch estimation based on RAPT framework // Signal Proc. Conf. (EUSIPCO). 2012. Proc. of the 20th European. 2012. P. 2787—2791.
5. **Abe T., Kobayashi T., Imai S.** Robust pitch estimation with harmonics enhancement in noisy environment based on instantaneous frequency // Proc. ICSLP96. 1996. V. 2. P. 1277—1280.
6. **Hasan M. K., Shahnaz C., Fattah S. A.** Determination of pitch of noisy speech using dominant harmonic frequency // Proc. IEEE Int. Symp. on Circuits and Systems. 2003. V. 2. P. 556—559.
7. **Cheveigne A., Kawahara H.** YIN, a fundamental frequency estimator for speech and music // J. Acoust. Soc. Am. 2002. V. 111(4). N. 4. P. 1917—1930.
8. **Yeh J. R., Shieh J. S., Huang N. E.** Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method // Advances in Adaptive Data Analysis. 2010. V. 2 (2). P. 135—156.
9. **Алимурадов А. К., Муртазов Ф. Ш.** Методы повышения эффективности распознавания речевых сигналов в системах голосового управления // Измерительная техника. 2015. № 10. С. 20—24.
10. **Alimuradov A. K., Churakov P. P.** Noise-robust speech signals processing for the voice control system based on the complementary ensemble empirical mode decomposition / Int. Siberian Conf. Control and Communications (SIBCON 2015). 2015. May 21—23. Russia. Omsk.
11. **Huang N. E., Zheng S., Steven R. L.** The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis // Proc. the Royal Society of London A. 1998. V. 454. P. 903—995.
12. **Zhaohua W., Huang N. E.** Ensemble empirical mode decomposition: A noise-assisted data analysis method // Advances in Adaptive Data Analysis. 2009. V. 1 (1). P. 1—41.
13. **Bhawna Sh., Sukhvinder K.** Distinction Between EMD & EEMD Algorithm for Pitch Detection in Speech Processing // Int. J. Engineering Trends and Technology (IJETT). 2014. V. 7. N. 3. P. 119—125.
14. **Schlotthauer G., Torres M. E., Rufiner H. L.** A new algorithm for instantaneous f0 speech extraction based on ensemble empirical mode decomposition / 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland. 2009. P. 2347—2351.
15. **Priyanka G., Mahendra Kumar P.** Determine the Pitch Markers in Speech Signal Using Ensemble Empirical Mode Decomposition, Int. J. Advanced Research in Computer Science and Software Engineering, 2012. V. 2 (7). N. 7. P. 90—96.
16. **Алимурадов А. К.** Исследование частотно-избирательных свойств методов декомпозиции на эмпирические моды для оценки частоты основного тона речевых сигналов // Труды МФТИ. 2015. Т. 7. № 3. С. 56—68.
17. **Huang X., Acero A., Hon H.-W.** Spoken Language Processing. Guide to Algorithms and System Developmen. Prentice Hall, Upper Saddle River, 2001.

Дата принятия: 09.02.2016.