УДК 004.5
DOI 10.18413/2411-3808-2019-46-2-359-366

СПОСОБ ВЫДЕЛЕНИЯ ТРАЕКТОРИИ ЧАСТОТЫ ОСНОВНОГО ТОНА РЕЧИ НА ОСНОВЕ ЧАСТОТНОЙ ДЕМОДУЛЯЦИИ

PITCH FREQUENCY SEPARATION METHOD BASED ON A FREQUENCY DEMODULATION

Д.Д. Полешенков, О.О. Басов
D.D. Poleshenkov, O.O. Basov

Федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики»,
Россия, 197101, Санкт-Петербург, Кронверкский пр., 49

Saint Petersburg National Research University of Information Technologies, Mechanics and Optics,
49 Kronverkskiy prospekt, St. Petersburg, 197101, Russia

E-mail: d.poleshenkov@yandex.ru, oobasov@mail.ru

Аннотация
В работе на основе экспериментальных данных сформулирован способ выделения траектории частоты основного тона, позволяющий упростить реализацию задачи классификации кадров речевого сигнала и обладающий более высокой точностью при равных затратах машинного времени. Данный способ построен на основе частотной демодуляции речевого сигнала в полосе частот основного тона. В ходе проведенных исследований выявлено, что результат работы выделителей основного тона для длительно произнесенных вокализованных фонем полностью совпадает с результатом частотной демодуляции речевого сигнала в полосе частот его первой гармоники (в полосе частот основного тона). Разработан алгоритм, реализующий указанный способ. Произведена первичная проверка работы алгоритма на практике. Показаны основные направления исследований по совершенствованию разработанного способа. Дополнительно приведено описание направлений практической реализации выделения траектории частоты основного тона на основе разработанного способа.

Abstract
This paper gives a pitch frequency separation method based on experimental data. This method allows simplifying a solution of a speech signal frames classification problem and increasing accuracy of pitch detection algorithms functioning. It has an increased accuracy and decreased computational complexity in comparison with the same algorithms. This method is based on a frequency demodulation process of a speech signal in a pitch band. The analysis of long-spoken vocal phonemes showed that the result of a pitch detector functioning is equal with the frequency demodulation result of the first harmonic band of a speech signal. The algorithm realizing this method was created. The description of proposed algorithm functioning is given in this paper. Primary examination of the proposed algorithm was made in this work. This paper gives the basic research ways of a proposed method improvement. In addition the description of a practical realization ways of a proposed pitch detection method based on frequency demodulation process was given in this paper.

Ключевые слова: речевой сигнал, кадр речевого сигнала, частота основного тона, фонемный переход, частотная демодуляция.
Keywords: speech signal, pitch frequency, frame, phoneme transition, frequency demodulation.

Введение

При решении задач, связанных с выделением траектории частоты основного тона (ОТ) в системах реального времени, актуальным является разработка способа оценки параметров ОТ с заданной точностью при минимальных затратах машинного времени. Анализ распространенных подходов к выделению траектории частоты ОТ и классификации кадров речевого сигнала (РС) [De Cheveigne, 2002; Лузин, 2009; Первушин, 2011; Жиляков, 2012; Басов, 2014; Вольф, 2015; Гапочкин, 2016; Вишнякова, 2016; Алимурадов, 2018] показал, что задача создания оптимального способа, реализующего указанные функции, далека от окончательного решения. Практическая реализация большинства рассмотренных способов является сложной и требует больших вычислительных ресурсов для получения достаточной точности. При этом задачи классификации кадров РС и выделения траектории частоты ОТ, как правило, рассматриваются обособленно и не учитывают некоторых структурных особенностей РС, позволяющих существенно уменьшить вычислительную сложность используемых алгоритмов.

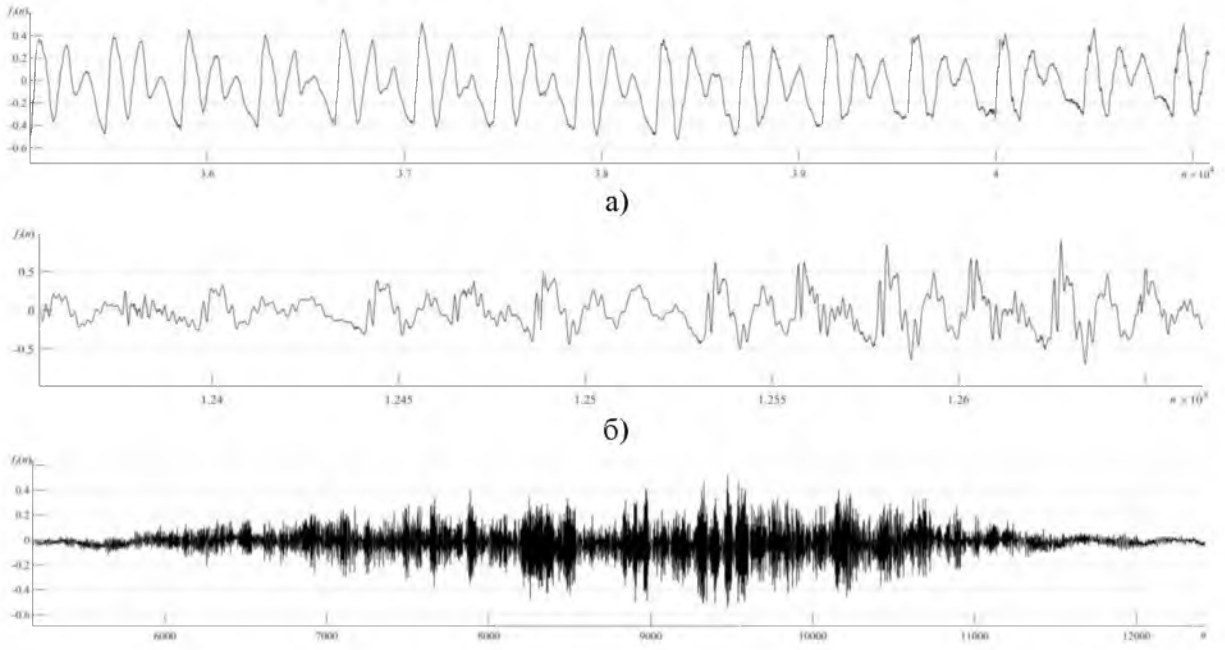
Исходя из проведенного анализа и полученных экспериментальных данных, представляется возможной разработка способа выделения траектории частоты ОТ, включающего в себя упрощенный механизм классификации кадров РС. Указанный способ должен учитывать структурные особенности синтеза вокализованных фонем и РС в целом, что позволит уменьшить вычислительную сложность с сохранением высокого качества оценки рассматриваемых параметров.

Описание структуры фонем и кадров РС

В связи с тем, что РС не является стационарным, возникает необходимость выделения и классификации однородных сегментов речи с целью оптимизации процесса обработки. Минимальной структурной единицей речи являются фонемы, которые по своей структуре и способу образования подразделяют на гласные, дифтонги, полугласные, носовые, фрикативные, взрывные, шумовые и аффрикаты [Rabiner, 1978]. Так как задача выделения фонемных переходов является сложной с точки зрения реализации и критериев определения [Fant, 1960], длительность кадра анализа, как правило, не совпадает с длительностью фонемы (для задач, не связанных с непосредственной обработкой фонем) и выбирается исходя из условия квазистационарности РС на выбранном временном интервале. Таким образом, в зависимости от фонемного набора на длительности сегмента анализа выделяют вокализованные, слабо вокализованные и невокализованные кадры РС (рис. 1). Существуют и другие подходы к формированию кадров анализа (например, [Сорокин, 2016]), однако они, как правило, требуют наличия дополнительных решающих устройств или усложняют алгоритм анализа.

При произнесении вокализованных участков речи источником сигнала возбуждения являются голосовые связки, квазипериодические осцилляции которых приводят к возникновению полигармонического волнового процесса. В резонансной системе речевого тракта происходит интерференция прямой и отраженных волн, что приводит к усилению или ослаблению сигнала соответствующих гармоник ОТ. При этом структура сигнала ОТ определяется физиологическим строением речевого аппарата и влиянием функционирования других систем организма и не зависит от текущих параметров резонансной системы.

Для слабо вокализованных участков речи характерно относительно плавное изменение параметров резонансной системы и значительное влияние шумовых компонент на структуру сигнала [Деркач, 1983], что приводит к существенному увеличению числа грубых ошибок оценки параметров ОТ. Существует достаточное количество способов минимизации числа указанных ошибок (например, [Бабкин, 2005; Леонов, 2017]), однако их практическая реализация приводит к значительному усложнению конечных приборов.



а)

б)

в)

Рис. 1. Кадры речевого сигнала (а – вокализованный; б – слабо вокализованный; в – шумоподобный)
Fig. 1. Speech signal frames (a – voiced; b – jittery voiced; c – unvoiced)

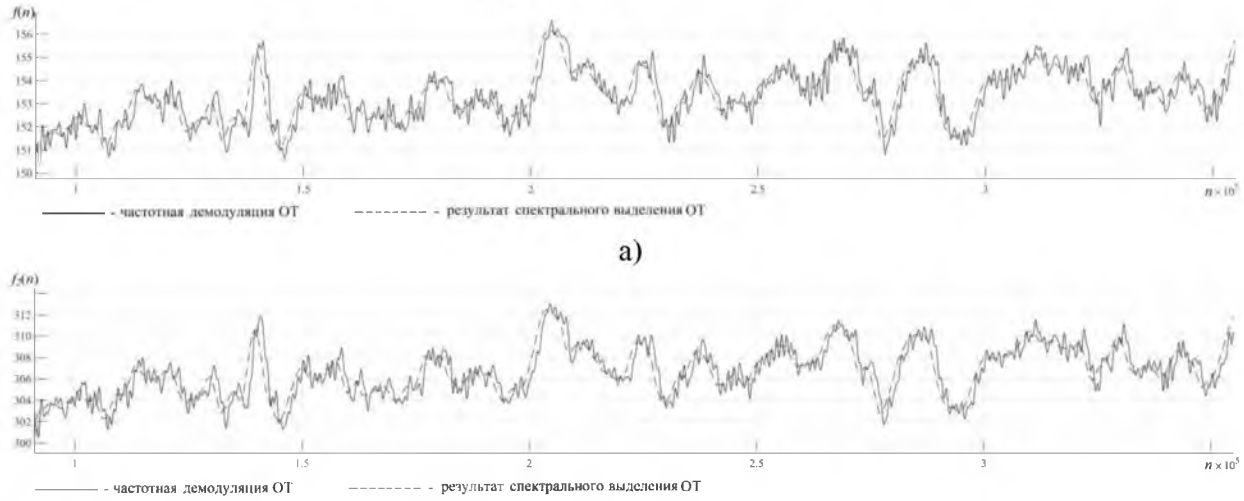
В связи с тем, что синтез шумоподобных речевых сегментов происходит при минимальном участии голосовых связок, достоверное выделение параметров ОТ на рассматриваемых интервалах РС не представляется возможным. Однако данная особенность позволяет достаточно точно идентифицировать шумоподобные фонемы на длительности РС.

Таким образом, с точки зрения выделения параметров ОТ, наибольший интерес представляет структура формирования вокализованных фонем в вокализованных и слабо вокализованных кадрах речевого сигнала, а именно зависимость параметров сигналов гармоник от изменения параметров сигнала ОТ.

Описание предлагаемого способа и алгоритма его реализации

Исходя из описания структуры резонансной системы речевого аппарата человека [Сорокин, 1992], можно сделать вывод о том, что в ней отсутствуют элементы, вносящие структурные нелинейные искажения, что приводит к ограничению возможности появления дополнительных спектральных составляющих РС (на вокализованных участках) кроме тех, что вызваны квазипериодическими осцилляциями голосовых связок. Как следует из модели колебаний голосовых связок [Сорокин, 1985], рассматриваемые осцилляции имеют фиксированный набор гармоник для заданной функции мышечного возбуждения. Следовательно, изменения параметров колебаний связок будут приводить к сильнокоррелированным откликам на частотах гармоник. Предполагается, что изменение функции мышечного возбуждения на фонемных переходах будет вносить дополнительные отличия в сигналы гармоник ОТ.

В ходе анализа длительно произнесенных вокализованных фонем ($t = 10...14$ с) было выявлено, что основной вклад в изменение частоты и амплитуды ОТ вносит функционирование сердечнососудистой системы, а результат работы алгоритмов выделения параметров ОТ для указанных сигналов полностью соответствует результату частотной демодуляции первой гармоники РС (рис. 2). Присутствие частотной модуляции в речи на вокализованных участках ранее описывалось в литературе [Леонов, 2009], однако причинами ее возникновения считались внутренние явления, возникающие при осцилляциях голосовых связок. Более глубокие экспериментальные исследования, направленные на установление степени зависимости траектории частоты ОТ от сигнала изменения давления выдыхаемого из легких воздуха, позволят уточнить существующие представления о данном процессе.



а)

б)

Рис. 2. Сравнение результатов работы спектрального выделителя ОТ с результатом частотной демодуляции ОТ для сигнала длительно произнесенной фонемы «а»
(а – для первой гармоники ОТ; б – для второй гармоники ОТ)
Fig. 2. Comparison of a spectral pitch detector functioning result with a frequency demodulation signal for long-spoken phoneme «a» (a – for the first harmonic; b – for the second harmonic)

Также установлено, что выделенные траектории частоты гармоник ОТ для сигналов вокализованных фонем можно описать выражением:

$$f_k(n) = f_{от}(n) \cdot k, \quad k = 1,2,3\ldots, \tag{1}$$

где $f_k(n)$ – сигнал траектории k-ой гармоники; $f_{от}(n)$ – сигнал траектории частоты ОТ; k – номер гармоники.

Косвенным подтверждением установленной зависимости может служить описанное в литературе (например, [Воробьев, 2006]) относительное постоянство разности фаз гармоник РС на его вокализованных сегментах.

Свойство, описанное выражением (1), может быть использовано в качестве классификационного признака вокализованных участков РС, однако для этого необходимо убедиться в том, что для шумоподобных фонем спектральные составляющие на кратных частотах значительно отличаются независимо от реализации сигнала фонемы.

Для проверки данной гипотезы был произведен анализ структуры шумоподобных фонем (объем выборки составил около 100 реализации каждой фонемы) посредством оценки коэффициента корреляции нормированных траекторий частоты спектральных составляющих на локальных максимумах спектра сигнала, находящихся в полосе частот ОТ, и спектральных составляющих на кратных частотах. Результаты проведенного анализа (таблица 1) позволяют судить о минимальной взаимосвязи между указанными компонентами сигналов фонем, что позволяет определить способ выделения траектории частоты ОТ путем сравнения траекторий частот его гармоник, выделенных посредством частотной демодуляции.

Таблица 1
Table 1

Коэффициенты корреляции спектральных составляющих для анализируемых фонем
Correlation coefficients for analyzed phonemes spectral components

Фонема	к	х	ф	п	ч	с	т	ш	щ	ц
$k_{кор} \cdot 10^{-3}$	-1,7	0,22	-2,2	2,4	-1,2	-0,27	3,1	-1,8	-2,0	-0,3

В алгоритме, реализующем указанный способ (рис. 3), используются следующие исходные данные: $s(n)$ – последовательность отсчетов исходного речевого сигнала; $f_д$ – частота дискретизации; p – значение порога принятия решения.

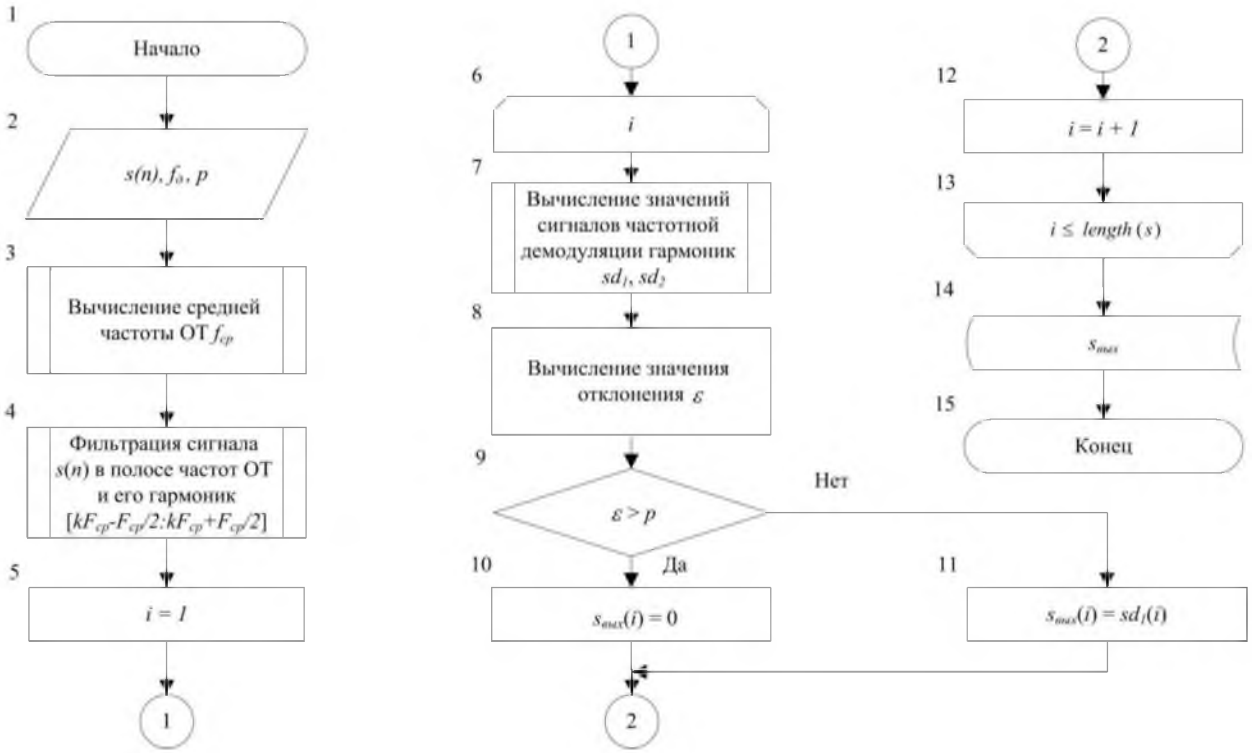


Рис. 3. Блок-схема алгоритма, реализующего предложенный способ
Fig. 3. Block diagram of the algorithm implementing the proposed method

При работе алгоритма после ввода данных происходит вычисление средней частоты ОТ ($f_{ср}$) (максимума спектра сигнала в полосе частот от 50 до 350 Гц, соответствующего ОТ) посредством дискретного преобразование Фурье. Далее происходит выделение полосы частот ОТ и предполагаемой полосы частот его первой гармоники, формируются соответствующие им сигналы $s_1(n)$ и $s_2(n)$, после чего осуществляется их частотная демодуляция с несущими $f_{ср}$ и $2f_{ср}$ соответственно. На основе результата демодуляции (сигналов $sd_1(n)$ и $sd_2(n)$ соответственно) происходит вычисление отклонения в соответствии с выражением:

$$\varepsilon(n) = \frac{\left| sd_1(n) - \frac{1}{2} sd_2(n) \right|}{f_{ср}} \qquad (2)$$

Далее осуществляется сравнение сигнала отклонения, вычисленного в соответствии с выражением (2), с пороговым значения p. Величина порога выбирается исходя из средней амплитуды изменения траектории частоты ОТ. При превышении отклонением порогового значения соответствующие отсчеты сигнала условно помечаются как шумоподобные и не учитываются при дальнейшем анализе. Таким образом, в результате работы алгоритма остаются полученные значения траектории частоты ОТ, с высокой вероятностью соответствующие вокализованным составляющим РС.

Оценка вычислительной сложности и проверка работы алгоритма

Так как в основе рассматриваемого алгоритма лежит процесс частотной демодуляции, а вычисление максимума в полосе частот ОТ выполняется единожды на длительности кадра, то вычислительную сложность рассмотренного алгоритма можно

оценить как O(N^2), при этом результат вычисления будет иметь максимальное разрешение по времени (одному отсчету временного представления РС соответствует один отсчет траектории частоты ОТ). Такое разрешение по времени могут обеспечить алгоритмы, основанные на корреляционном и спектральном методе оценивания параметров ОТ. Вычислительная сложность данных алгоритмов при той же глубине вычислений оценивается как O(N^3) и O($N^2 log(N)$) соответственно. Стоит отметить, что для решения задач оценки траектории частоты ОТ, где не требуется точное определение абсолютного значения частоты, реализацию предложенного способа можно осуществить на основе аналоговых схемотехнических решений с использованием фазовой автоподстройки частоты [Баскаков, 2000].

Результаты работы алгоритма (рис. 4), построенного на основе описанного способа, в целом позволяют сделать вывод о его работоспособности. В приведенном примере на интервале произнесения звука «а» траектории частоты ОТ совпадают, при этом проявляется эффект сглаживания сигнала результата спектрального выделения ОТ, обусловленный значительной длительностью кадра анализа. Однако на интервалах, соответствующих звукам [л] и [н'], наблюдается значительное количество отброшенных значений. Это обусловлено большой полосой пропускания фильтра сигнала ОТ, вследствие чего сказывается влияние артикуляционной составляющей указанных звуков. Предполагается, что использование фазовой автоподстройки частоты в совокупности с уменьшением полосы пропускания фильтра ОТ позволит минимизировать подобные эффекты.

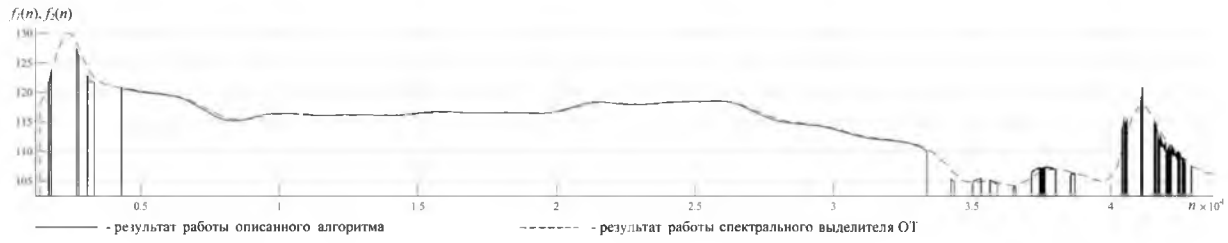


Рис. 4. Сравнение результатов работы алгоритма, построенного на основе предлагаемого способа, с результатами работы спектрального алгоритма выделения траектории частоты ОТ для записи слова «лань»

Fig. 4. Comparison of the algorithm functioning results based on the proposed method with the spectral algorithm functioning results for the word «lan'»

Важным условием для корректности получаемых оценок является отсутствие помех (например, шума промышленной сети электропитания и его гармоник) в полосе частот ОТ анализируемого РС. Минимизация влияния рассматриваемых негативных эффектов не входит в задачи описанного алгоритма, так как он предназначен для оценки возможности реализации предложенного способа вычисления траектории частоты ОТ.

Заключение

Применение полученных результатов в задачах проектирования систем обработки речи позволяет оптимизировать вычислительные затраты. Дальнейшее совершенствование предложенного способа выделения траектории частоты основного тона, связанное с определением критериев классификации выделяемых фонем и минимизацией негативного влияния шума и помех, позволит повысить качество существующих алгоритмов кодирования, распознавания и синтеза речи.

Работа выполнена при финансовой поддержке фонда РФФИ (проект № 18-07-00380).

Список литературы
References

1. Алимурадов А.К., Квитка Ю.С., Чураков П.П., Тычков А.Ю. 2018. Повышение точности измерения частоты основного тона на основе оптимизации процесса декомпозиции речевых сигналов на эмпирические моды. Измерение. Мониторинг. Управление. Контроль. 4(26): 53–65.

Alimuradov A.K., Kvitka Yu.S., Churakov P.P., Tychkov A.Yu. 2018. Increasing the accuracy of measuring the pitch frequency based on the optimization of the process of decomposition of speech signals on empirical modes. Measuring. Monitoring. Management. Control. 4(26): 53–65.

2. Баскаков С.И. 2000. Радиотехнические цепи и сигналы. М., Высшая школа, 462.

Baskakov S.I. 2000. Radiotechnicheskie cepi i signaly [Radiotechnical circuits and signals]. Moscow, Vysshaya shkola, 462.

3. Басов О.О., Носов М.В., Шалагинов В.А. 2014. Исследование характеристик джиттера периода основного тона речевого сигнала. Труды СПИИРАН. 1(32): 27–44.

Basov O.O., Nosov M.V., Shalaginov V.A. 2014. Pitch-jitter analysis of the speech signal. SPIIRAS Proceedings. 1(32): 27–44.

4. Бабкин В.В. 2005. Помехоустойчивый выделитель основного тона речи. Цифровая обработка сигналов и ее применение: материалы 7-й международной конференции М., ИПУ РАН. Доклады, X-1: 175–178.

Babkin V.V. 2005. [Noise-immune extractor of the speech pitch]. Cifrovaja obrabotka signalov i ee primenenie: materialy 7-j mezhdunarodnoj konferencii [Digital signal processing and its application: proceedings of the 7th international conference]. M., IPU RAN, Doklady, X-1: 175–178.

5. Вишнякова О.А, Лавров Д.Н. 2016. Гибридный алгоритм выделения частоты основного тона. Математические структуры и моделирование. Омск, Омский государственный университет. 1(37): 59–65.

Vishnyakjova O.A., Lavrov D.N. 2016. The hybrid algorithm of extraction of fundamental frequency. Mathematical structures and modeling. 1(37): 59–65.

6. Воробьев В.И, Давыдов Г.В., Шамгин Ю.В. 2006. Фазовые соотношения между основным тоном и обертонами гласных звуков. Доклады БГУИР. 2(14): 64–68.

Varabyeu V.I., Davydau G.U., Shamgin YU.V. 2006. Phase relation between fundamental tones and vowel sounds obertones. Doklady BSUIR. 2(14): 64–68.

7. Вольф Д.А., Мещеряков Р.В. 2015. Модель и программная реализация сингулярного оценивания частоты основного тона речевого сигнала. Труды СПИИРАН. 6(43): 191–209.

Volf D.A., Meshcheryakov R.V. 2015. Software Implementation of a Singular Meter of the Pitch Frequency of a Speech Signal. SPIIRAS Proceedings. 6(43): 191–209.

8. Гапочкин А.В. 2016. Определение основного тона речи с помощью вейвлет-преобразования и его применение. Вестник МГУИП имени Ивана Федорова. Москва. Московский государственный университет имени Ивана Федорова. 1: 22–24.

Gapochkin A.V. 2016. Determination of the main tone of speech using the wavelet transform and its application. Vestnik MGUP imeni Ivana Fedorova. Moscow. Moscow state university of print n.a. Ivan Fedorov. 1: 22–24.

9. Деркач М.Ф., Гумецкий Р.Я. 1983. Динамические спектры речевых сигналов. Львов, Вища школа, 168.

Derkach M.F., Gumetskiy R.Ya. 1983. Dinamicheskie spektry rechevikh signalov [Dinamic spectrums of speech gignals]. L'vov, Vischa shkola [L'vov, High school], 168.

10. Жиляков Е.Г., Фирсова А.А., Чеканов Н.А. 2012. Алгоритмы обнаружения основного тона речевых сигналов. Научные ведомости Белгородского государственного университета. Серия «Экономика. Информатика». 1(120): 135–143.

Zhilyakov E.G., Firsova A.A., Chekanov N.A. 2012. Detection algorithm of the fundamental tone speech signals. Belgorod State University Scientific Bulletin. Economics Information technologies. 1(120): 135–143.

11. Лузин Д.А. 2009. Разработка и исследование системы автоматического выделения основного тона речи. Автореф. дис. … тех. наук. Ижевск, 26.

Luzin D.A. 2009. Razrabotka i issledovanie sistemy avtomaticheskogo vydeleniya osnovnogo tona rechi [Design and research of automatic pitch detection]. Abstract. dis. … cand. tech. sciences. Izhevsk, 26.

12. Первушин Е.А., Лавров Д.Н. 2011. Алгоритм выделения основного тона и детектирования тон/не тон по минимумам разностной функции на участке минимального периода. Математические структуры и моделирование. Омск, Омский государственный университет. 1(22): 24–27.

Pervushin E.A., Lavrov D.N. 2011. Algoritm vydeleniya osnovnogo tona rechi detectirovaniya ton/ ne ton po minimumam raznostnoi' funktcii na uchastke minimal'nogo perioda [Pitch extraction algorithm and tone/not tone detection according to minimum of difference function on the segment of minimal period]. Mathematical structures and modeling. 1(22): 24–27.

13. Сорокин В.Н. 1985. Теория речеобразования. М., Радио и связь, 312.

Sorokin V.N. 1985. Teoriya recheobrazovaniya [Speech synthesis theory]. Moscow, Radio i svyaz', 312.

14. Сорокин В.Н. 1992. Синтез речи. М., Наука, 392.

Sorokin V.N. 1992. Sintez rechi [Speech synthesis]. Moscow, Nauka, 392.

15. De Cheveigné A., Kawahara H. 2002. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America. 111(4): 1917–1930.

16. Fant G. 1960. Acoustic theory of speech production, Hague, Mouton, 328

17. Leonov A.S., Makarov I.S., Sorokin V.N. 2009. Frequency modulations in the speech signal. Acoustical physics. 55(6): 876–887.

18. Leonov A.S., Sorokin V.N. 2017. Upper bound of errors in solving the inverse problem of identifying a voice source. Acoustical physics. 63(5): 570–582.

19. Rabiner, L.R., Schafer, R.W. 1978. Digital Processing of Speech Signals. Englewood Cliffs, NJ. Prentice-Hall, 512.

20. Sorokin V.N. 2016. Segmentation of the period of the fundamental tone of a voice source. Acoustical physics. 62(2): 244–254.