**DOI: 10.15587/2312-8372.2018.134977**

# РАЗРАБОТКА МЕТОДА АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ПОЛА ДИКТОРА НА ОСНОВЕ СОВМЕСТНОГО ОЦЕНИВАНИЯ МОМЕНТОВ ЧАСТОТЫ ОСНОВНОГО ТОНА И ФОРМАНТНЫХ ЧАСТОТ

**Омельченко С. В.**

## 1. Введение

Алгоритмы распознавания пола диктора необходимы для решения ряда прикладных задач. Результаты определения пола диктора используются в системах адаптивного распознавания слов и фонем речи, идентификации и верификации дикторов, поскольку распознавание пола диктора позволяет существенно сузить область принимаемых признаками значений.

Размеры гортани, голосовых складок и мышц, управляющих их колебаниями, различны у мужчин и женщин. Это дает основания для поиска различительных признаков в параметрах импульсов голосового возбуждения и цифрового фильтра модели речеобразования.

Поэтому актуальным является исследование методов распознавания пола диктора по речевым сигналам.

## 2. Объект исследования и его технологический аудит

*Объект исследования* – методы распознавания пола диктора по речевым сигналам.

Одним из наиболее важных этапов, который, в конечном счете, будет определять качество классификации, является выбор классификационных признаков.

Как правило, в качестве информационного параметра, по которому проводится идентификация пола диктора, используют частоту основного тона. Однако, как показывает практика, одной частоты тона недостаточно для достоверной классификации пола диктора.

Другими характерными недостатками, которые присущи данному объекту в существующих условиях функционирования, является сложность реализации и низкая устойчивость в условиях действия помех высокого уровня.

## 3. Цель и задачи исследования

*Целью работы* является разработка алгоритмов автоматического распознавания пола диктора.

Для достижения цели были поставлены следующие задачи:

1. Выбрать новые классификационные признаки.

2. Разработать построение решающего правила (классификатора), который являются устойчивыми к действию помех, гаусовский белый шум.

3. Провести экспериментальные исследования разработанных алгоритмов.

## 4. Исследование существующих решений проблемы

Предыдущие исследования по гендерной идентификации предложили множество особенностей и методов классификации. Выделение функции часто

выполняется с использованием гендерных характеристик речи, таких как частота тона, которую дополняют кепстральными признаками [1, 2]. Другие подходы основаны на спектральных особенностях, таких как коэффициенты линейного предсказания, коэффициенты отражения. Методы классификации используют скрытые марковские модели, гауссовские модели смеси [3, 4]. Также разработаны многоэксплуатационные подходы, сочетающие методы классификации.

В работе [5] рассмотрены особенности распознавания пола диктора по 4-м формантным частотам и 12 Мел-кепстральным коэффициентам (MFCC), где получена вероятность правильного распознавания 0,94.

В существующей системе [6] используются нечеткая логика и нейронные сети. Однако такая система не дает высокого качества гендерной классификации и сложна в реализации из-за сложности обучения сети.

В частности, работа [7] посвящена рассмотрению особенности распознавания пола по речи, полученного с телефона. Рассмотрены различные методы классификации, включая метод k-ближайшего соседа, Байесовский подход, многослойный персептрон с использованием в качестве признаков Мел-кепстральных коэффициентов (MFCC). При этом получена вероятность правильного распознавания – 0,90.

В [8] рассмотрены особенности распознавания пола по частоте основного тона и Мел-кепстральным коэффициентам (MFCC) использовании логистической и линейной регрессии. При этом получена вероятность правильного распознавания – 0,95.

В системе [9] строятся гауссовские смеси для Мел-кепстральных коэффициентов (MFCC). Такая система с 24 коэффициентами MFCC и 16 компонентами гауссовской смеси распределений имеет до 100 % правильного распознавания. Однако такая система сложна в реализации и обучении.

В альтернативном варианте решения проблемы, изложенной в [10], строятся гауссовские смеси для Мел-частотных коэффициентов (MFCC). Такая система имеет 92 % правильного распознавания.

Результаты анализа позволяют сделать вывод о том, что алгоритмы распознавания пола диктора, как правило, сложны в реализации и не удовлетворяют по качеству распознавания пола диктора.

**5. Методы исследования**

Полагается, что на вход системы распознавания поступает временная последовательность отсчетов речевого сигнала $s(n)$, $n=\overline{0,N-1}$, взятых с интервалом дискретизации $\Delta t$.

Необходимо построить алгоритм, который по предъявленной реализации речи выносит решения о принадлежности текущих структурных речевых единиц к заданным типам, классам и обеспечивал бы максимум средней вероятности правильного распознавания пола дикторов $P_{pr}$.

Рассмотрим работу распознавателя пола диктора. С целью получения динамических признаков распознаваемого цифрового сигнала производится

разбиение слов на отрезки одинаковой длительности, которая обычно составляет 10–30 мс.

Вначале для составления хранимых эталонов речевых единиц диктора выполняется сегментация слов, фонем. Подобная сегментация на этапе распознавания речевых единиц позволяет исключить избыточные процедуры принятия решений по сигналам, не несущим речевую информацию либо не являющимися целостными речевыми единицами. Задачи сегментации состоит в членении речи на структурные единицы и оценивании их временных границ. Алгоритмы сегментации подробно рассмотрены в [11, 12].

Полагая, что в пределах выборки речевой сигнал стационарен в широком смысле, алгоритм фильтрации речевого сигнала в частотной области имеет вид:

$$\hat{x}(t) = \mathrm{Re}\left( (N)^{-1/2} \sum_{m=0}^{N-1} C(m) H_{кор}(m) \exp\left( i\left( \frac{2\pi t}{N} \right) m \right) \right),$$
$$C(m) = (2N)^{-1/2} \sum_{\tau=0}^{2N-1} y_\tau^j \exp\left( -i\left( \frac{2\pi\tau m}{2N} \right) \right), \quad (1)$$

где $y_i^j = \begin{cases} s_i^j, & i = 0,1,...,(N-1), \\ 0, & i = N,(N+1),...,(2N-1) \end{cases}$ – входные отсчеты; $H_{кор}(m)$ – амплитудно-частотная характеристика фильтра.

Одним из основных параметров устной речи является частота повторения колебаний голосовых связок при произнесении вокализированной речи, называемая «Основным тоном» (ОТ). Для распознавания можно использовать особенности распределения частоты основного тона. Измерения по речевым сигналам, выполненные для голосов пяти дикторов-женщин показали, что диапазоны возможных значений частоты основного тона от 135 до 522, а для пяти дикторов-мужчин от 58 до 238 Гц. Хотя диапазоны оценок частот основного тона для мужчин и женщин перекрываются, но различаются средними частотами основного тона 128 Гц для мужчин и 256 Гц для женщин.

Для оценки частоты основного тона лучше использовать блоки, которые являются вокализованными. Существуют множество методов вычисления признак вокализованности. Так например, признак вокализованности $N_j^u$ вычисляется путем подсчета количества нуль-пересечений для каждой из выборок *j*-й выборки *u*-го сегмента слова. Принятие решения о вокализованности производится по сравнению с порогом, вычисленным, например, гистограммным методом.

Гистограммы несимметричны – относительно своей моды: у женских голосов (рис. 1), со стороны малых периодов склон круче, чем для больших периодов, тогда как у мужчин наблюдается обратная картина (рис. 2). Для таких распределений адекватными будут гамма распределения. Можно для распознавания использовать куммулянты до 6-го порядка, в том числе и нечетных.

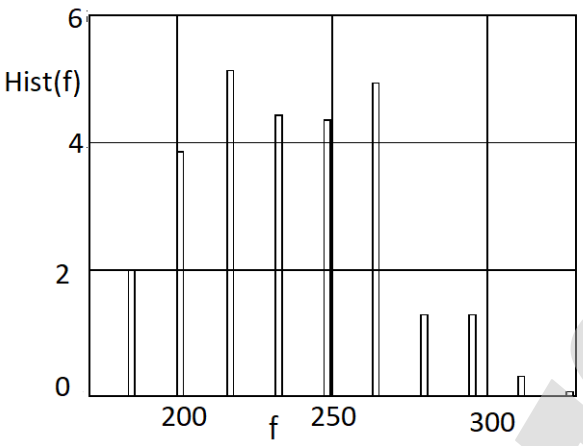


**Рис. 1.** Гистограмма частоты основного тона для женского голоса

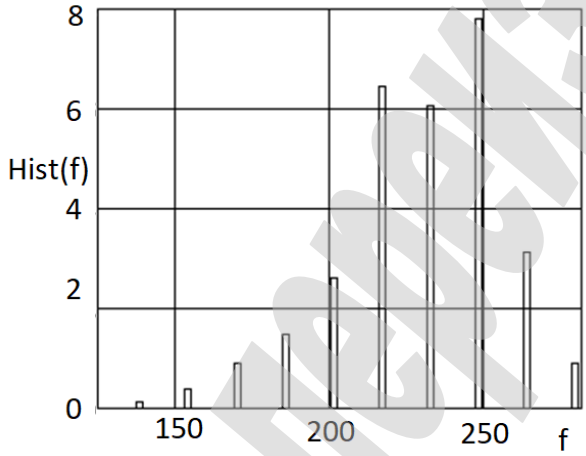


**Рис. 2.** Гистограмма частоты основного тона для мужского голоса

Кумулянтный коэффициент $\gamma_k$ определяется следующим образом:

$$\gamma_k = \frac{\mu_k}{\mu_2^{k/2}}, \tag{2}$$

где $\mu_k$ – кумулянты порядка $k$, однозначно связанные с центральными моментами.

Вычисление асимметрии и эксцесса позволяет установить симметричность распределения случайной величины $X$ относительно математического ожидания $M(x)$. Для этого находят третий центральный момент, характеризующий асимметрию закона распределения случайной величины. Если он равен нулю $\mu_3 = 0$, то случайная величина $X$ симметрично распределена относительно математического ожидания $M(X)$. Поскольку $\mu_3$ имеет размерность случайной величины в кубе, то вводят безразмерную величину – коэффициент асимметрии:

$$As = \frac{\mu_3}{\sigma^3}. \tag{3}$$

Центральный момент четвертого порядка используется для определения эксцесса, характеризует плосковершинность или островершинность плотности вероятности. Эксцесс вычисляется по формуле:

$$Es = \frac{\mu_4}{\sigma^4} - 3. \tag{4}$$

Оценка центрального момента по частоте основного тона $f_o$:

$$\mu_i = \sum_{k=1}^{N} \left( f_{ok} - M(f_o) \right)^i. \tag{5}$$

Из рис. 3 видно, что можно провести разделяющую линейную границу между женскими и мужскими классами.

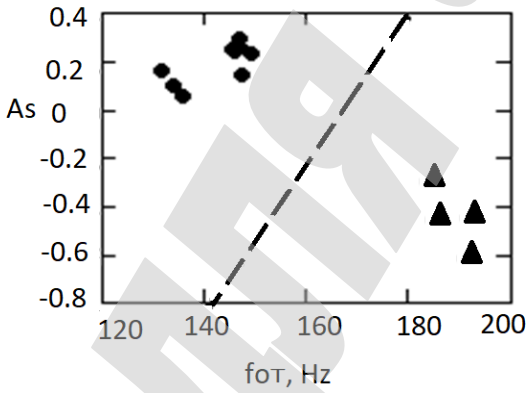


**Рис. 3.** Сосредоточение пар измерений среднего значения частоты основного тона $f_{от}$ и асимметрии *As* для женских (треугольники) и мужских голосов (точки)

Алгоритм принятия решений:

$$i = \mathrm{sgn}\left( mf_{от} - k_1 \cdot As - fz \right), \tag{6}$$

где $\mathrm{sgn}(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$ – знаковая функция; $mf_{от}$ – оценка среднего значения частоты основного тона; $As$ – оценка коэффициента асимметрии; $k_1$, $f_z$ – коэффициенты решающего правила.

Учитывая экспериментально полученное сосредоточение пар измерений среднего значения частоты основного тона $f_{от}$ и асимметрии *As* для женских и мужских голосов (рис. 3), получим примерные значения коэффициентов решающего правила *fr*=170 и $k_1$=178.

Исследования показали, что использование оценок дисперсии $D(x) = \sigma^2$ и эксцесса $Es$ не позволяют использовать четкую разделяющую границу, хотя их использование возможно.

Можно использовать дополнительно и другие признаки, улучшающие распознавание пола диктора и учитывающие особенности голосового тракта

Анализ влияния изменения длины речеобразующего тракта на параметры голоса показал, что уменьшение длины речеобразующего тракта приводит к существенному росту частот формант. Это объясняет наличие более высоких частот формант в женском голосе по сравнению с мужским.

Существуют методы оценивания формантных частот на основе оценок коэффициентов авторегрессии, кепстральных признаков [13–15].

Оценки формантных частот спектрально-полосным методом [16, 17] для каждого из блоков могут вычисляться как среднеэффективные частоты с соответствующего выхода полосового фильтра. В табл. 1 для каждого *m*-го фильтра указанны граничные частоты $f_в^{(m)}$ и $f_н^{(m)}$.

**Таблица 1**

Граничные частоты фильтров

| *m* | $f_н^{(m)}$, Гц | $f_в^{(m)}$, Гц |
|---|---|---|
| 1 | 200 | 850 |
| 2 | 850 | 2200 |
| 3 | 2200 | 3000 |
| 4 | 3000 | 4000 |

Рассмотрим особенности формирования формантно-полосных признаков. Из совокупности отсчетов формируются блоки (выборки) речи, которые берутся с 2–3-х кратным перекрытием или без него. Согласно этому методу вычисляют спектрально-полосные сигналы, соответствующие вероятному расположению формант, полосы которых приведены в табл. 1. Граничные частоты $f_в^{(m)}$, $f_н^{(m)}$ соответствуют *m*-м формантам при частоте дискретизации 8 кГц.

При этом оценки формантных частот для заданной выборки вычисляются путем подсчета количества нуль-пересечений речевого сигнала с соответствующего выхода полосового фильтра.

Процедура вычисления формант может быть повторена, но при этом в качестве граничных полос частот используют:

$$\hat{f}_в^{(m)} = \hat{f}^{(m)} + \Delta,\ \hat{f}_н^{(m)} = \hat{f}^{(m)} - \Delta,$$

где $\hat{f}^{(m)}$ – форманты, вычисленные на предыдущем этапе; $\Delta$ – границы диапазона поиска формант. Простейшей среди рекуррентных процедур является двухэтапная.

Исследования показали, что наибольший вклад в распознавание вносит первая и вторая форманта и показатель их эксцесса, где возможно линейное разделения классов. На рис. 4 показано полученное экспериментальное сосредоточение пар измерений среднего значения второй формантной частоты и коэффициента эксцесса $E_a$ для пяти женских (треугольники) и пяти мужских

голосов (квадраты). Это позволяет использовать для второй форманты линейные разделяющие границы между ними. Из проведенных экспериментальных получено, что подобное линейное разделение возможно и для первой форманты и ее коэффициента эксцесса, а для третьей и четвертой формант затруднительно.

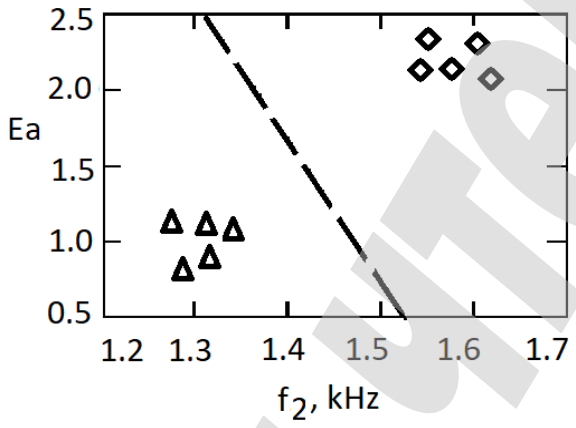


**Рис. 4.** Сосредоточение пар измерений среднего значения второй формантной частоты $f_2$ и коэффициента эксцесса $E_a$ для женских (треугольники) и мужских голосов (квадраты)

Алгоритм принятия решений:

$$i = \mathrm{sgn}\left( mf_{om} - k_0 \cdot As - fz + \sum_{k=1}^{4} k_k \left( mf_k - f_{zp_k} \right) + \sum_{k=1}^{2} ke\left( Es_k - Es_{zpk} \right) \right), \qquad (7)$$

где $mf_{om}$ – оценка среднего значения частоты основного тона; $As$ – асимметрия частоты основного тона; $mf_k$ – оценка значения частот формант; $Es_k$ – оценка коэффициента эксцесса для частоты $k$-ой форманты.

**6. Результаты исследования**

Испытания приведенных выше алгоритмов распознавания слов проводились на основе данных, веденных в ЭВМ с микрофона через звуковой интерфейс с частотой дискретизации $F_д$=8 кГц.

Испытания проводились на реальных выборках звуковых сигналов, введенных в ЭВМ с выхода микрофона.

Экспериментальные исследования алгоритмов распознавания слов речи с одноэтапным определением количества нулей в полосах формант проводились методом статистического испытаний на выборках 10-и сигналов для каждого из 10-х различных дикторов мужского и женского пола. По выборкам оценивались параметры решающего правила, а также контрольные выборки реальных сигналов использовались для оценивания качества распознавания сигналов.

Алгоритм принятия решений, с учетом ранее рассмотренного правила принятия решений (7), имеет вид:

$$i = \mathrm{sgn}\left( \frac{f_{om}}{178} - As - \frac{f_2}{1400} - \frac{Es_2}{1{,}5} + 1 \right), \quad (8)$$

где $f_{om}$ – оценка среднего значения частоты основного тона; As – оценка асимметрии частоты основного тона; $f_2$ – оценка значения частоты 2-й форманты; $Es_2$ – оценка эксцесса для частоты 2-ой форманты.

Алгоритм принятия решений по оценке среднего значения и коэффициента асимметрии частоты основного тона:

$$i = \mathrm{sgn}\left( \frac{f_{om}}{178} - As - 1 \right). \quad (9)$$

Алгоритм принятия решений по оценке среднего значения и эксцесса частоты 2-ой форманты:

$$i = \mathrm{sgn}\left( 2 - \frac{f_2}{1400} - \frac{Es_2}{1{,}5} \right). \quad (10)$$

По результатам статистических испытаний для каждого из алгоритмов (8)–(10) получена оценка средней вероятности правильного распознавания $P_{pr}$=1. При дополнительном действии аддитивной помехи типа гауссов белый шум и отношении сигнал/шум $q$=20 получены экспериментально оценки средних вероятностей правильного распознавания $P_{pr}$. Для алгоритма принятия решений в соответствии с формулой (9) экспериментально получена оценка средней вероятностей правильного распознавания – $P_{pr}$=0,9. Для алгоритма принятия решений в соответствии с формулой (8) – $P_{pr}$=0,8, а для алгоритма, принятия решений в соответствии с формулой (7) – $P_{pr}$=0,7. Таким образом, при действии аддитивной помехи типа гауссов белый шум высокого уровня рационально использовать алгоритм принятия решений в соответствии с формулой (9), который учитывает особенности оценок среднего значения частоты основного тона и оценок асимметрии частоты основного тона.

Проведенные исследования подтверждают эффективность приложенных алгоритмов.

**7. SWOT-анализ результатов исследования**

*Strengths*. По сравнению с аналогами, положительное действие объекта исследований в виде составных элементов системы распознавания заключается в оптимизации выбора признаков принятия решений с целью повышения вероятности правильного распознавания пола диктора в зависимости от уровня шума. Это включает моделирование системы распознавания на персональном компьютере.

*Weaknesses.* К слабым сторонам предлагаемых эффективных параметров системы распознавания можно отнести необходимость начальных капитальных вложений в систему распознавания пола. Также необходимо предусмотреть расходы на их производство по месту использования. Также к слабым сторонам предлагаемых решений можно отнести их локальность («точечность») по отношению ко всей комплексной системе распознавания для различных языков.

*Opportunities.* Предлагаемые технические решения по повышению качества распознавания способствует улучшить качество распознавания дикторов, распознавания слов речи, упростить поиск в базах данных. Это, в свою очередь, позволит существенно уменьшить затраты на производство систем. Ожидаемая прибыль прогнозируется получить примерно через 2–3 года в зависимости от количества систем.

В перспективе предполагается использование полученных результатов не только для русского и украинского языков, но и для ряда иностранных языков.

*Threats.* От предприятия или эксплуатирующей организации потребуются начальные капитальные вложения в технической реализации системы распознавания. Также необходимы затраты на их производство. Отрицательное воздействие на объект исследования внешних факторов в виде внешней среды и других условий эксплуатации обусловлены нормативным сроком эксплуатации. Это зависит от используемых разработок. Однако этот срок составляет не менее 5 лет, что является более чем достаточным, для самоокупаемости разработанных организационно-технических решений.

**8. Выводы**

1. Выбраны новые классификационных признаки, включающие совместное использование оценок среднего значения частоты основного тона, её коэффициента эксцесса, оценок средних значений формант и их коэффициентов асимметрии.

2. Построены решающие правила принятия решений о поле диктора на основе линейного разделения по взвешенной сумме оценок предложенных классификационных признаков. Линейные границы для разделения полов дикторов обусловлены компактным расположением признаков для каждого из типов дикторов.

3. По найденным рабочим характеристикам проведены сравнительные исследования алгоритмов распознавания пола диктора.

По результатам статистических испытаний для алгоритма включающие совместное использование оценок среднего значения частоты основного тона, её коэффициента эксцесса, оценок средних значений формант и их коэффициентов асимметрии получена оценка средней вероятности правильного распознавания 1. При дополнительном действии аддитивной помехи типа гауссов белый шум и отношении сигнал/шум $q$=20. Для такого алгоритма экспериментально получена вероятность правильного распознавания – 0,8. Для алгоритма принятия решений, использующего лишь оценки среднего значения частоты основного тона и её коэффициента эксцесса, получена оценка средней

вероятности правильного распознавания – 0,9. Это говорит о большей помехоустойчивости таких алгоритмов.

Проведенные исследования алгоритмов распознавания подтверждают возможность получения приемлемого качества распознавания пола диктора на основе использования:

- оценок среднего значения частоты основного тона и её коэффициентов асимметрии;
- оценок среднего значения частот формант;
- оценок коэффициентов эксцесса для формантных частот.

**Литература**


1. Kalyuzhnyi A. Ya., Semenov V. Yu. Metod identifikatsii pola diktora na osnove modelirovaniya akusticheskikh parametrov golosa gaussovymi smesyami // Akustichniy visnik. 2009. Vol. 12, No. 2. P. 31–38.
2. Scheme E., Castillo-Guerra E., Englehart K., Kizhanatham A. Practical Considerations for Real-Time Implementation of Speech-Based Gender Detection // Lecture notes in computer science. 2006. Vol. 4225. P. 426–436. doi: http://doi.org/10.1007/11892755_44
3. Sorokin V. N., Makarov I. S. Opredelenie pola diktora po golosu // Akusticheskiy zhurnal. 2008. Vol. 54, No. 4. P. 659–668.
4. Robust GMM-based gender classification using pitch and RASTA-PLP parameters of speech / Zeng Y.-M. et al. // Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. Dalian, 2006. P. 3376–3379. doi: http://doi.org/10.1109/icmlc.2006.258497
5. Faek F. Objective Gender and Age Recognition from Speech Sentences // Aro, The Scientific Journal of Koya University. 2015. Vol. 3, No. 2. P. 24–29. doi: http://doi.org/10.14500/aro.10072
6. Jayasankar T., Vinothkumar K., Vijayaselvi A. Automatic Gender Identification in Speech Recognition by Genetic Algorithm // Applied Mathematics & Information Sciences. 2017. Vol. 11, No. 3. P. 907–913. doi: http://doi.org/10.18576/amis/110331
7. Gender Identification using MFCC for Telephone Applications – A Comparative Study / Ahmad J. et al. // International Journal of Computer Science and Electronics Engineering. 2015. Vol. 3, No. 5. P. 351–355.
8. Levitan S. I., Mishra T., Bangalore S. Automatic identification of gender from speech // Proceeding of Speech Prosody. 2016. P. 84–88. doi: http://doi.org/10.21437/speechprosody.2016-18
9. Yucesoy E., Nabiyev V. V. Gender identification of a speaker using MFCC and GMM // 2013 8th International Conference on Electrical and Electronics Engineering (ELECO). Bursa, 2013. doi: http://doi.org/10.1109/eleco.2013.6713922
10. Harb H., Chen L. Gender identification using a general audio classifier // 2003 International Conference on Multimedia and Expo. ICME ’03. Proceedings (Cat. No.03TH8698). Baltimore, 2003. doi: http://doi.org/10.1109/icme.2003.1221721
11. Presnyakov I. N., Omelchenko S. V. Pomekhoustoychivye algoritmy segmentatsii rechi v sistemakh obrabotki // Radiotekhnika. 2003. No. 131. P. 165–177.

12. Sorokin V. N., Tsyplikhin A. I. Segmentatsiya i raspoznavanie glasnykh // Informatsionnye protsessy. 2004. Vol. 4, No. 2. P. 202–220.

13. Presnyakov I. N., Omelchenko A. V., Omelchenko S. V. Avtomaticheskoe raspoznavanie rechi kanalakh peredachi // Radioelektronika i informatika nauchno-tekhnicheskiy zhurnal. 2002. No. 1. P. 26–31.

14. Rabiner L. R., Schafer R. W. Digital Processing of Speech Signals. Pearson; US edition, 1978. 962 p.

15. Marple S. L. Digital Spectral Analysis: With Applications/Disk,Pc/MS Dos/IBM/Pc/at. Prentice Hall Signal Processing Series, 1987. 492 p.

16. Presnyakov I. N., Omelchenko S. V. Avtomaticheskoe raspoznavanie razdel'nykh slov i fonem rechi // Radioelektronika i informatika. 2003. No. 2. P. 41–47.

17. Presnyakov I. N., Omelchenko S. V. Algoritmy raspoznavaniya rechi // Avtomatizirovannye sistemy upravleniya i pribory avtomatiki. 2004. No. 126. P. 136–145.