УДК 684.3

В. И. Соловьев
Восточноукраинский национальный университет
Молодежный квартал, 20-А, 91034 Луганск, Украина

Идентификация характеристик голоса на основе максимумов вейвлет-преобразования

Приведены результаты исследований по разработке методологии и математической модели выявления и анализа самоподобных структур в речевых фрагментах аудиофайлов. Модель позволяет на основе вейвлетного базиса Морле выделять в пространственной области скейлограмм геометрические образования типа «хребтов». Показано, что эти образования при определенных условиях имеют взаимно однозначное соответствие с индивидуальными характеристиками голоса.

Ключевые слова: *самоподобные структуры, фрагменты речи, идентификация индивидуальных характеристик голоса.*

Научно-обоснованный подход к задаче идентификации индивидуальных характеристик голоса насчитывает около 150 лет. Он базируется на основных концепциях теории речеобразования, основы которой заложил Гельмгольц [1]. За прошедшее время многочисленные исследования в различных областях знания, изучающих вопросы акустики генерации и восприятия звука человеком, в целом, фактически подтвердили основные положения теории речеобразования [2–16]. В то же время, множество задач обработки речевой информации, таких например, как автоматическая идентификация фонемических составляющих речи, идентификация характеристик голоса по характеристикам аудиоинформации трудно поддаются эффективному решению. На сегодняшний день человек, по-прежнему, решает эти задачи гораздо эффективнее созданных автоматических систем.

Основной сегодняшней проблемой в этой группе задач является выявление эффективных подходов, которые позволяют выделить в речевой информации устойчивые характеристики, связанные с фонемическими составляющими и индивидуальными характеристиками голоса.

В предложенной статье рассматривается задача автоматической идентификации индивидуальных характеристик голоса по характеристикам временного ряда амплитуды звуковой волны речевого файла. В последнее десятилетие было предложено множество подходов и моделей для ее эффективного решения. Однако попрежнему отсутствует достаточная ясность в вопросе — какие именно характеристики звукового аудиофайла ответственны за индивидуальные характеристики го-

лоса. Нет ясности также в ответе на вопрос — какими математическими методами и преобразованиями можно получить устойчивые эффективные оценки этих характеристик.

Ниже приводятся результаты разработки модели, показавшей при практической реализации высокую эффективность решения данной задачи.

Постановка задачи исследования

Рассмотрим задачу, характерную для криминалистической экспертизы аудиоданных. Предположим, имеется один или несколько цифровых аудиофайлов с записью речевой информации. При этом голоса принадлежат различным физическим лицам. Необходимо на основе математического анализа статистических характеристик речевых фрагментов файла или нескольких файлов выявить речевые фрагменты, характеристики голоса в которых идентичны.

Будем рассматривать цифровые аудиофайлы, записанные в формате wav. Проанализируем цифровые аудиоданные как временные ряды амплитуды звуковой волны. Нашей задачей в дальнейшем является выделение статистических характеристик тех параметров аудиофайлов, которые можно было бы с определенной вероятностью интерпретировать как характеристики, обусловленные индивидуальными физическими характеристиками голоса человека.

Очевидно, что при любой автоматической идентификации характеристик голоса важную роль играет эффективная автоматическая сегментация аудиоданных на паузы и фрагменты речи. Данная задача решалась на основе мультифрактального подхода [17–22] и здесь не рассматривается.

Рассмотрим фрагменты речи без пауз как дискретные временные ряды амплитуды звуковой волны A_i (i — номер временного отсчета).

При визуальном рассмотрении графика изменения амплитуды звуковой волны фрагмента речи (рис. 1) можно выделить множество геометрически самоподобных структур на различных временных масштабах.

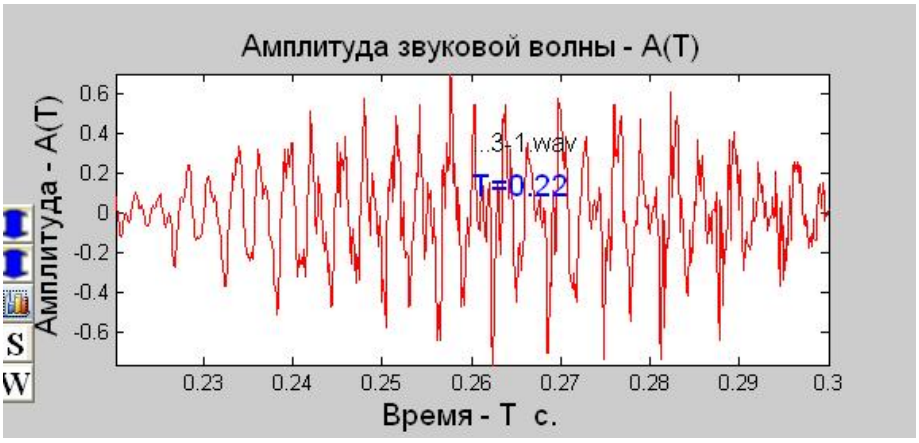

Рис. 1. Иллюстрация графика изменения звуковой волны фрагмента речи

В дальнейшем геометрически самоподобные структуры различных речевых фрагментов аудиофайла будем рассматривать как мультифрактальные структуры,

в соответствии с концепцией Мандельброда [18–21]. Можно предположить, что индивидуальные характеристики голоса должны быть связаны с определенными самоподобными мультифрактальными структурами на определенных временных масштабах звуковой волны. Фактически любой подход при решении задач идентификации характеристик голоса прямо или косвенно должен выявлять и анализировать мультифрактальные структуры, связанные с индивидуальными характеристиками голоса. Так, например, любое построение спектра амплитуды звуковой волны на основе оконного преобразования Фурье позволяет оценить наиболее существенные составляющие спектра частот. На основе этой информации возможно построение различных статистических характеристик, полезных при идентификации индивидуальных характеристик голоса. Однако эффективность автоматических систем, построенных на основе таких подходов, весьма низкая.

Далее рассмотрим задачу идентификации индивидуальных характеристик голоса на основе анализа характеристик сигналов голоса. Важным вопросом при исследовании любых частотных и иных характеристик речевых сигналов является вопрос выбора временного окна (интервала) при их анализе. Любые характеристики будут существенно отличаться при рассмотрении в различных временных масштабах.

Дальнейший анализ будем проводить на временных масштабах не более 20–30 мс. На этом временном интервале еще не сформированы фонемические характеристики речи. С другой стороны, на интервалах порядка 20 мс уже достаточно явно проявляются индивидуальные особенности характеристик голосовых сигналов [2].

Существенным фактором спектрального анализа амплитуды звуковой волны является вид вейвлет-преобразования, которое применяется для исследования. Степень сродства вейвлетного базиса и фрагментов речевого файла очевидным образом весьма существенно влияет на эффективность исследования.

В дальнейшем исследовании используем вейвлет-базис Морле [23]. Поставим задачу выявления и эффективного сравнения характеристик мультифрактальных структур в речевых файлах, которые возможно являются следствием индивидуальных характеристик голоса.

Модель выявления самоподобных структур на основе максимумов вейвлет-преобразования

Рассмотрим фрагменты речи в аудиоданных, как дискретный временной ряд амплитуды звуковой волны. Поставим задачу выделения характеристик самоподобных структур в полученном временном ряду. Это могут быть самые различные геометрически подобные структуры, используемые при рассмотрении графиков изменения амплитуды звуковой волны. Будем, в соответствии с развиваемой концепцией, рассматривать самоподобие, как геометрическое подобие, связанное с преобразованиями сжатия, растяжения — как по оси времени, так и по амплитудной координате. Для выявления подобных структур используем методы вейвлет-анализа [22]. С этой целью выберем вейвлет-базис, имеющий существенное сродство с геометрическими структурами амплитуды звуковой волны. Из всего многообразия базисов, используемых на сегодняшний день в практике вейвлет-

анализа, наиболее близок по геометрическим структурам к динамике амплитуды звуковой волны в аудиофайлах комплексный вейвлет Морле [23]

$$C_{mor}(t_i, T_k, F_b, F_c) = (\pi F_b)^{0,5} \exp(2j\pi F_c t_i) \exp(-(t_i - T_k)^2 / F_b), \quad (1)$$

где F_b — параметр ширины вейвлета; F_c — центральная частота вейвлета; t_i — дискретные временные отсчеты; T_k — временной отсчет соответствующий центральной части временного окна; j — комплексная единица.

Пусть $A(t_i)$ — значение амплитуды звуковой волны фрагмента речи аудиофайла в момент времени t_i. Рассмотрим временное окно фрагмента речи с интервалом δT. Этот интервал, как указано выше, в исследованиях принимался в диапазоне 10–30 мс. Параметр ширины комплексного вейвлета Морле F_b выберем постоянным для всех преобразований. Его величина выбиралась из условия практического затухания абсолютных значений вейвлета Морле при значениях $t_i - T_k$, равных δT /2. Для каждого фрагмента речи вычислим свертку вейвлета Морле с фрагментом временного ряда амплитуды звуковой волны в виде

$$C(T_k, F_b, F_c) = (1/N) abs\left(\sum_{t_{ij}=0}^{N_m} C_{mor}(t_i, T_k, F_b, F_c) A(t_i) \right), \quad (2)$$

где $C(T_k, F_b, F_c)$ — значение модуля коэффициента вейвлет-преобразования; N — количество дискретных отсчетов на интервале δT временного окна.

При фиксированном параметре ширины F_b комплексного вейвлета Морле значение модуля является функцией частоты F_c вейвлета Морле и положения временного окна во времени — T_k. Типичный график пространственной скейлограммы $C(T_k, F_b, F_c)$ в функции F_c и T_k представлен на рис. 2.

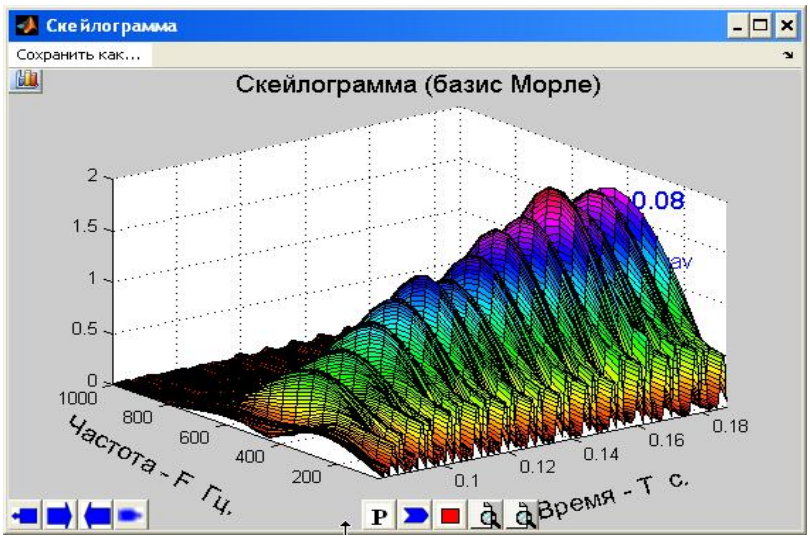

Рис. 2. Скейлограмма фрагмента речи

Рассматриваемое в рамках разработанной модели представление фрагмента речи в виде пространственной скейлограммы обладает рядом важных особенностей, позволяющих существенно повысить эффективность выявления самоподобных структур. В частности, локальные максимумы вейвлет-преобразования являются весьма информативными с точки зрения выявления и анализа самоподобных структур, связанных с индивидуальными характеристиками голоса. Эти структуры, как будет показано далее, являются весьма стабильными образованиями в скейлограммах.

Модель идентификации индивидуальных характеристик голоса на основе максимумов вейвлет-преобразования

На рис. 3 представлена характерная иллюстрация, на которой совмещены фрагменты аудиозаписи речи и скейлограмма, построенная на основе рассматриваемой модели.

Анализ показывает, что расположение «хребтов» скейлограмм по временному параметру на рис. 3 строго соответствует локальным экстремумам амплитуды звуковой волны во временной области. Причем эти локальные экстремумы соответствуют всплескам амплитуды звуковой волны, обусловленными частотой основного тона. Но наиболее существенной особенностью характеристик «хребтов» является форма «хребта». Исследования показывают, что после рациональной нормировки функции $C(T_k, F_b, F_c)$ при фиксированных T_k и F_b на вершине «хребта» эти функции обладают высокой степенью геометрического подобия. При этом форма нормированных «хребтов» индивидуально отлична при отличии характеристик голоса.

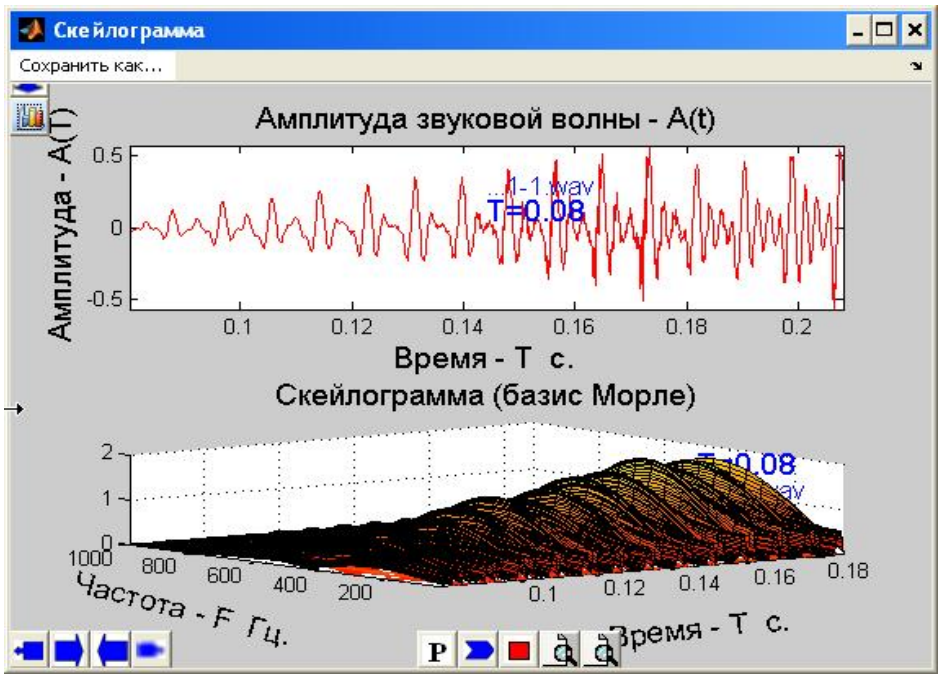

Рис. 3. Иллюстрация выявления самоподобных структур

Причиной высокой степени подобия структур типа «хребет» в частотной области на скейлограмме является специфика сродства вейвлета Морле временной структуре амплитуды звуковой волны в областях локальных экстремумов. Вейвлет-преобразование Морле при данном подходе эффективно выделяет эти структуры. Рассматриваемое подобие имеет весьма прозрачную физическую трактовку. Структуры амплитуды звуковой волны в области локальных экстремумов, соответствующих частоте основного тона, имеют достаточно явно выраженную геометрическую симметрию относительно амплитуды локального экстремума. При этом вейвлет Морле в силу сродства позволяет выявить эту симметрию в виде явно выраженных экстремумов скейлограмм.

Важным фактором вейвлет-анализа в рассматриваемом подходе, определяющим эффективность решения рассматриваемых задач, является высокое частотно-временное разрешение. Во времени это разрешение максимально возможное. По порядку величины оно соответствует минимальному интервалу дискретизации.

Применение в разрабатываемом подходе вейвлетного базиса Морле является одним из самых существенных факторов, обеспечивающих эффективное выделение самоподобных структур. Так, например, на рис. 4 приведена иллюстрация трехмерного спектра Фурье того же фрагмента речи, что и на рис. 1. Трехмерный спектр построен с использованием оконной функции Ханна [23].

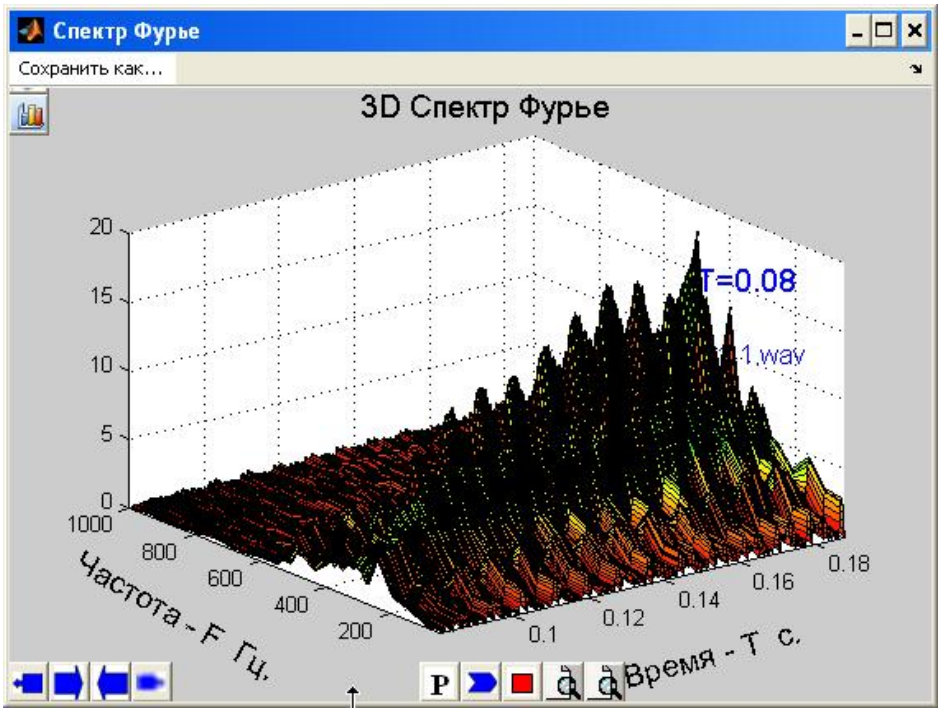

Рис. 4. Трехмерный спектр Фурье фрагмента речи с окном Ханна

На рис. 4 визуально легко выделить самоподобные фрагменты. Однако выделение этих пространственных самоподобных структур математическими методами представляет собой по-прежнему сложную задачу, как в частотной, так и во

временной областях. Выделение трехмерных структур после вейвлет-преобразования Морле является сравнительно легко реализуемой задачей.

В разрабатываемой модели структурами, соответствующими индивидуальным характеристикам голоса, являются пространственные частотно-временные структуры, которые в дальнейшем будем называть «хребтами».

Спецификой этих структур в некотором диапазоне параметров F_b является математическая «гладкость» функции $C(T_k, F_b, F_c)$ по параметрам T_k и F_c. Эта «гладкость» не является строгой. Но она весьма существенна с точки зрения эффективного выделения пространственных структур «хребтов» для дальнейшего анализа. Экспериментальные исследования показали, что индивидуально, каждый из «хребтов» может быть выделен путем определения и исключения областей локальных минимумов функции $C(T_k, F_b, F_c)$ по временной переменной T_k при фиксированном параметре частоты F_c.

После выделения «хребты» нормируются к единице:

$$C_m(T_k, F_b, F_c) = C(T_k, F_b, F_c) / \sum_{T_k, F_c} C(T_k, F_b, F_c). \qquad (3)$$

Здесь $C_m(T_k, F_b, F_c)$ — нормированная скейлограмма индивидуального «хребта». Суммирование в формуле (3) распространяется лишь на отсчеты выделенного индивидуального «хребта». После такой нормировки функции $C_m(T_k, F_b, F_c)$ возможно трактовать как трехмерные плотности распределения вероятностей. Для сравнения нормированных плотностей распределения вероятностей, с целью принятия или отвержения гипотезы об идентичности распределения вероятностей, использовалась модифицированная версия критерия Колмогорова-Смирнова [24]. Детальное рассмотрение модификации в данной статье не излагается (будет рассмотрено в дополнительной статье).

Массовые экспериментальные исследования многочисленных речевых файлов и различных фрагментов речи на основе разработанной модели показали следующее. Для одного и того же индивидуального голоса наблюдается весьма высокая степень подобия геометрии структур «хребтов». При этом существенным параметром подобия, который необходимо учитывать в статистических критериях Колмогорова-Смирнова, является глобальный максимум индивидуального «хребта» по частоте F_c. Если индивидуальный набор «хребтов» сгруппировать по значению глобального максимума частоты, то ошибка первого рода принятия гипотезы об идентичности пространственных распределений при невысоком уровне шумов, как правило, менее 0,05.

При отличии величины глобального максимума для «хребтов» для проверки подобия осуществляется смещение частот «хребтов», и проверяются на идентичность модифицированные плотности распределения вероятностей. При этом для одних и тех же индивидуальных характеристик голоса эти распределения, как показывают критерии согласия, по-прежнему идентичны.

Второй вариант сравнения «хребтов» моделирует выявление одного и того же голоса в аудиофайле при наличии нескольких говорящих. Смена характеристик

голоса в файле практически мгновенно идентифицируется, как отвержение гипотезы об идентичности плотностей распределения вероятностей. Для этого на новом фрагменте речи необходим интервал времени порядка нескольких десятков миллисекунд.

Разработанная модель позволяет решать большинство актуальных задач идентификации индивидуальных характеристик голоса.

Выводы

Разработана модель выявления самоподобных структур в речевых аудиофайлах на основе вейвлетного базиса Морле. Модель позволяет установить взаимно однозначное соответствие между индивидуальными характеристиками голосовых сигналов и пространственными структурами скейлограмм типа «хребтов». Разработана модель выявления и сравнения пространственной геометрии подобных «хребтов». Подход позволяет решать множество актуальных задач идентификации индивидуальных характеристик голосовых сигналов.

1. Helmholtz H. von. Die Lehe von Tonempfindungen / Helmholtz H. Von. — Brannschweig, Vieweg, 1863.

2. *Психоакустические* аспекты восприятия речи. Механизмы деятельности мозга: под. ред. Н.П. Бехтеревой. — М.: Наука, 1988. — 504 с.

3. *Галунов В.И.* Помехоустойчивость как системообразующий фактор речи. Проблемы и методы экспериментальных исследований / В.И. Галунов. — СПб., 2002. — С. 295.

4. *Фланаган Дж.* Анализ, синтез и восприятие речи / Дж. Фланаган: пер. с англ.; под ред. А.А. Пирогова. — М.: Связь, 1968. — 396 с.

5. *Фант Гуннар.* Анализ и синтез речи / Фант Гуннар. пер. с англ. В.С. Лозовского и Н.В. Бахмутовой; под ред. Н.Г. Загоруйко. — Новосибирск.: Наука. Сиб. отд., 1970. — 167 с.

6. *Цвикер Э.*Ухо как приемник информации / Э. Цвикер, Р. Фельдкеллер: пер. с нем. под ред. Б.Г. Белкина. — М.: Связь, 1971. — 225 с.

7. *Вокодерная* телефония. Методы и проблемы: под ред. А.А. Пирогова. — М.: Связь, 1974. — 536 с.

8. *Бодуэн де Куртенэ И.А.* Разница между фонетикой и психофонетикой / Бодуэн де Куртенэ И.А. // Избранные труды по общему языкознанию. — 1963. — Т. 2. — С. 547.

9. *Ликлайдер Дж. К.Р.* Механические свойства слуха / Дж. К.Р. Ликлайдер, Дж. Розенблит // В кн. Экспериментальная психология: под ред. С.С. Стивенса, ИИЛ. — М., 1963. — Т. 2. — 1035 с.

10. *Винцюк Т.К.* Анализ, распознавание и интерпретация речевых сигналов / Т.К. Винцюк. — К.: Наук. думка, 1987. — 262 с.

11. *Найда С.А.* Объективная аудиометрия на основе формулы среднего уха — новый метод исследования и дифференциальной диагностики слуха / С.А. Найда // Электроника и связь. — 2004. — № 23. — С. 66–70.

12. *Алдошина И.А.* Основы психоакустики / И.А. Алдошина //Звукорежиссер. — 2000. — № 6. — С. 36–40.

13. *Сорокин В.Н.* Теория речеобразования / В.Н. Сорокин. — М.: Радио и связь, 1985. — 312 с.

14. *Вемян Г.В.* Передача речи по цепям электросвязи / Г.В. Вемян. — М. Радио и связь, 1985. — 272 с.

15. *Phase* in speech pictures / A.V. Oppenheim, J.S. Lim, G.E. Corec and S.C. Pohlig // Proc. IEEE Int. Conf Acoust. Sprech and Signal Processing. — 1979, Apr. — P. 632–637.

16. *Аграновский А.В.* Теоритические аспекты алгоритмов обработки и классификации речевых сигналов / А.В. Аграновский, Д.А. Леднов. — М.: Радио и связь. — 434 с.

17. *Федер Е.* Фракталы / Е. Федер. — М.: Мир, 1991. — 326 с.

18. *Mandelbrot B.* Statistical Methodology for Non-Periodic Cycles: From the Covariance to R/S Analysis / B. Mandelbrot. — Annals of Economic Social Measurement 1, 1972.

19. *Mandelbrot B.* The Fractal Geometry of Nature / B. Mandelbrot. — New York: W.H. Freeman, 1982.

20. *Mandelbrot B.* A Multifractal Walk Down Wall Street / B. Mandelbrot. — Scientific American, 1999.

21. *Mandelbrot B.B.* Robustness of the Rescaled Range R/S in the Measurement of Non-Cycling Long-Run Statistical Dependence / B.B. Mandelbrot // Water Resources Research. — 1969. — Vol. 5. — P. 967–988.

22. *Павлов А.Н.* Мультифрактальный анализ сложных сигналов / А.Н. Павлов, В.С. Анищенко // Успехи физических наук. — 2007. — Т. 177, № 8.

23. *Малла С.* Вейвлеты в обработке сигналов / С. Малла. — М.: Мир, 2005. — 670 с.

24. *Орлов А.И.* Теория принятия решений / А.И. Орлов: учебник. — М.: Экзамен, 2006. — 573 с.

Поступила в редакцию 03.05.2013