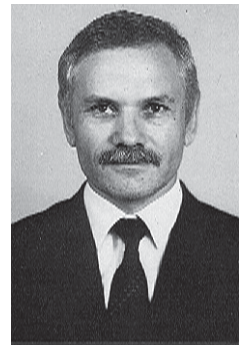
Александр Владимирович Брешенков родился в 1955 г., окончил МВТУ им. Н.Э. Баумана в 1982 г. Д-р. техн. наук, доцент кафедры “Компьютерные системы, комплексы и сети” МГТУ им. Н.Э. Баумана. Автор 70 научных работ в области САПР ЭВМ и баз данных.

A.V. Breshenkov (b. 1955) graduated from the Bauman Moscow Higher Technical School in 1982. Ph. D. (Eng.), assoc. professor of "Computer Systems, Complexes and Networks" department of the Bauman Moscow State Technical University. Author of 70 publications in the field of systems of automated design and data bases.

Александр Викторович Балдин родился в 1951 г., окончил МВТУ им. Н.Э. Баумана в 1974 г. Д-р техн. наук, начальник отдела интеграции информационных систем МГТУ им. Н.Э. Баумана. Автор 73 научных работ в области автоматизации и моделирования процессов управления и баз данных.



A.V. Baldin (b. 1951) graduated from the Bauman Moscow Higher Technical School in 1974. D. Sc. (Eng.), head of department for integration of information systems of the Bauman Moscow State Technical University. Author of 73 publications in the field of automation and simulation of management processes and data bases.

УДК 621.391:681.317

В. Б. Кропотов, Н. В. Медведев,
И. И. Троицкий

АНАЛИЗ СПОСОБОВ ИЗВЛЕЧЕНИЯ ХАРАКТЕРИСТИЧЕСКИХ ПРИЗНАКОВ РЕЧИ С ИСПОЛЬЗОВАНИЕМ ВЕЙВЛЕТОВ ПРИ РЕШЕНИИ ЗАДАЧИ РАСПОЗНАВАНИЯ ГОЛОСА ДИКТОРА В УСЛОВИЯХ СЛОЖНОЙ ШУМОВОЙ ОБСТАНОВКИ

Рассмотрены основные понятия преобразований сигналов (Фурье, Вейвлет), а также исследованы основные модели распознавания речи на основе этих преобразований. В процессе анализа алгоритмов распознавания предложенных моделей выявлены основные недостатки и преимущества каждой из моделей, а также даны рекомендации по использованию в конкретных условиях зашумления.

Выделение характеристических признаков говорящего человека – основа систем распознавания голоса диктора. При этом использование “сырого” сигнала без предварительной обработки практически не дает положительного результата. Классическим методом при анализе дискретных сигналов является быстрое преобразование Фурье (БПФ) с окном. Однако при анализе сигнала в зашумленной обстановке в

качестве основы для извлечения характеристических признаков говорящего человека из речевого сигнала использование вейвлет базисов представляется более эффективным. В настоящей работе рассмотрены три варианта использования вейвлет-преобразований в модуле извлечения характеристических признаков речи системы распознавания голоса диктора. Проведено сравнение производительности подобных систем с классическими системами, использующими преобразование Фурье.

Основные понятия. *Преобразование Фурье* управляет линейной инвариантной во времени обработкой сигнала f, так как синусоидальные волны e^{iwt} — это собственные функции инвариантных во времени операторов. Линейный инвариантный во времени оператор L полностью определяется собственным числом $\hat{h}(w)$:

$$\forall \omega \in R, \quad Le^{iwt} = \hat{h}(w)e^{iwt}. \tag{1}$$

Чтобы вычислить Lf, представляем сигнал f в виде суммы синусоидальных собственных функций $\{e^{iwt}\}_{w \in R}$:

$$f(t) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \hat{f}(\omega)e^{iwt}d\omega. \tag{2}$$

Если f обладает конечной энергией, то амплитуда $\hat{f}(\omega)$ каждой синусоидальной волны есть преобразование Фурье f:

$$\hat{f}(\omega) = \int\limits_{-\infty}^{+\infty} \hat{f}(t)e^{-iwt}dt. \tag{3}$$

Применяя оператор L к f в формуле (2) и подставляя выражение (1) для собственной функции, получаем

$$Lf(t) = \frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \hat{f}(\omega)\hat{h}(\omega)e^{iwt}d\omega. \tag{4}$$

Оператор L увеличивает или уменьшает каждую синусоидальную компоненту e^{iwt} функции f на множитель $\hat{h}(\omega)$. В этом и состоит частотная фильтрация f.

Принцип неопределенности устанавливает, что энергетическая протяженность функции и преобразование Фурье не могут быть одновременно малыми. В работе [1] элементарные частотно-временные атомы определены как волновые образования, имеющие минимальную протяженность на частотно-временно́й плоскости. Разложение сигнала по этим элементарным атомным образованиям позволяет получить содержание частотно-временно́й “информации” в анализируемом сигнале. Атомы Габора построены с помощью сдвига по

времени и частоте временного окна g:

$$g_{u,\xi}(t) = g(t-u)e^{i\xi t}. \tag{5}$$

Энергия $g_{u,\xi}$ сосредоточена в окрестности u на интервале размера σ_t, измеряемого стандартным отклонением $|g|^2$. Ее преобразование Фурье есть сдвиг на ξ преобразования Фурье $\hat{g}$ функции:

$$\hat{g}_{u,\xi}(\omega) = \hat{g}(\omega-\xi)e^{-iu(\omega-\xi)}. \tag{6}$$

Поэтому энергия $\hat{g}_{u,\xi}$ локализована около частоты ξ на интервале размера σ_ω. В частотно-временно́й плоскости (t,ω) протяженность энергии атома $g_{u,\xi}$ символически представляется прямоугольником Гейзенберга с центром в точке (u,ξ), который имеет временную ширину σ_t и частотную ширину σ_ω. Согласно принципу неопределенности можно утверждать, что площадь прямоугольника Гейзенберга удовлетворяет неравенству $\sigma_t\sigma_\omega \geqslant \dfrac{1}{2}$.

Эта площадь минимальна, когда g — функция Гаусса, в этом случае $g_{u,\xi}$ называют функциями Габора.

Преобразование Фурье с окном коррелирует сигнал f с каждым атомом $g_{u,\xi}$

$$Sf(u,\xi) = \int\limits_{-\infty}^{+\infty} f(t)g(t-u)e^{-i\xi t}dt. \tag{7}$$

Такое преобразование называется кратковременным преобразованием Фурье, потому что умножение на $g(t-u)$ локализует интеграл Фурье в окрестности $t=u$. Его дискретный аналог называется быстрым преобразованием Фурье с окном и может быть записан в следующем виде:

$$Sf[m,l] = \sum_{n=0}^{N-1} f[n]g[n-m]\exp\frac{-il2\pi n}{N}. \tag{8}$$

Предположим, что для любой точки (u,ξ) существует единственный атом $\phi_\gamma(u,\xi)$ с центром в точке (u,ξ) в частотно-временно́й плоскости. Частотно-временной прямоугольник для $\phi_\gamma(u,\xi)$ — это окрестность (u,ξ), где энергию f можно определить как

$$P_T f(u,\xi) = \left| \int\limits_{-\infty}^{+\infty} f(t)\phi^*(u,\xi)dt \right|^2. \tag{9}$$

Плотность энергии, называемую спектрограммой, можно найти по формуле [2]:

$$P_S f(u,\xi) = |Sf(u,\xi)|^2 = \left| \int\limits_{-\infty}^{+\infty} f(t)g(t-u)e^{-i\xi t}dt \right|^2. \tag{10}$$

По спектрограмме измеряют энергию f в частотно-временно́й окрестности, определяемой прямоугольником Гейзенберга для $g_{u,\xi}$.

Вейвлет-преобразование. Применение вейвлетов в задачах обработки и распознавания голоса продиктовано особенностями речевого акустического сигнала. Вейвлеты как средство многомасштабного анализа позволяют выделять одновременно основные характеристики сигнала и короткоживущие высокочастотные явления в речевом сигнале. Это свойство является существенным преимуществом вейвлетов в задачах обработки речевого сигнала по сравнению с оконным преобразованием Фурье с окном, где, изменяя ширину окна, приходится выбирать масштаб явлений, которые необходимо выделить в сигнале.

Вейвлет ψ — это функция с нулевым средним значением

$$\int\limits_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{11}$$

и параметрами сдвига u и растяжения s, имеющая вид

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right). \tag{12}$$

Вейвлет-преобразование f с масштабом s и сдвигом u вычисляется корреляцией f с вейвлет-атомом:

$$Wf(u,s) = \int\limits_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)dt, \tag{13}$$

где ψ^* — комплексно-сопряженное $\psi \in C$.

Как и преобразование Фурье с окном, применяя вейвлет-преобразование, можно определять частотно-временны́е изменения спектральных компонент, но вейвлет-преобразование имеет другое частотно-временно́е разрешение.

Фильтрация сигнала от шума. Зашумленный сигнал $X[n]$ может быть представлен в виде суммы $X[n] = f[n] + W[n]$, где $f[n]$ — полезный сигнал, а $W[n]$ — шум. Применительно к задаче распознавания речи диктора, отметим, что $f[n]$ — это голос диктора, а $W[n]$ — окружающая обстановка, оказывающая негативное влияние на качество работы системы распознавания голоса диктора.

Задача фильтрации состоит в сохранении компоненты $f[n]$ и подавлении шумовой составляющей $W[n]$. Вейвлет-преобразование позволяет проводить высококачественное разделение речевого сигнала на компоненты и его фильтрацию от шума.

Классическая модель системы распознавания голоса диктора. В общем случае система распознавания голоса диктора состоит из нескольких модулей. Базовым блоком является модуль извлечения индивидуальных особенностей голоса диктора. В большинстве современных систем распознавания для извлечения особенностей голоса используется БПФ с окном (8) в качестве основы.

Схематично работу системы можно представить следующим образом (рис. 1).

Аналоговый сигнал с наложенным на него шумом окружающей обстановки поступает на вход аналогово-цифрового преобразователя, после чего над получаемым дискретным сигналом выполняется БПФ с окном. В результате этого преобразования получается спектрограмма. Далее на стадии обучения извлекаются характеристические признаки говорящего человека и после обобщения полученных признаков для голоса каждого диктора строится эталонная модель. На основе имеющейся информации происходит оценка допустимых порогов классификации. Во время нормальной работы системы эталонные модели используются для принятия решения о принадлежности характеристических признаков конкретному диктору.

Подобные схемы показывают достаточно хорошую производительность в идеализированном окружении, но при применении в специфических условиях, например в зашумленной окружающей обстановке, качество их работы снижается [3].

Модель системы распознавания голоса диктора с разделенными задачами фильтрации и распознавания. За последние годы произошел значительный рост производительности вычислительных устройств, что позволяет строить системы распознавания голоса диктора на основе новых базисов разложения сигнала, которые, в свою очередь, позволяют существенно уменьшить влияние окружающей обстановки на производительность системы и качество работы системы в целом, для чего необходима фильтрация входного

Рис. 1. Классическая модель системы распознавания голоса диктора

сигнала перед последующей обработкой. Известные на сегодняшний день методы вейвлет-преобразования позволяют значительно уменьшить уровень шума в исходном сигнале. На рис. 2 приведена система с модулем фильтрации сигнала, основанная на дискретном вейвлет-преобразовании (ДВП).

Отметим следующие преимущества предложенной схемы:

— повышение качества распознавания голоса диктора в зашумленной окружающей обстановке;

— возможность использования системы распознавания как основной составляющей, которая базируется на классической модели.

Аналог предложенного метода был использован в работе [4] для решения задачи распознавания речи в зашумленной окружающей обстановке и позволил увеличить процент распознавания на 0,7 % в идеализированной обстановке без шума и на 28 % в сильно зашумленной обстановке.

Недостатком является тот факт, что не используется дополнительная информация, которую можно получить из сигнала с помощью вейвлет-преобразования.

Модель системы распознавания голоса диктора с совмещенными задачами фильтрации и распознавания. Применение вейвлет-преобразования позволяет выполнять анализ сигнала сразу на нескольких уровнях. При фильтрации шума с помощью ДВП исходный сигнал раскладывается по вейвлет-базису и имеется возможность анализа

Рис. 2. Модель системы распознавания голоса диктора с разделенными (*а*) и совмещенными (*б*) задачами фильтрации и распознавания

этого сигнала на нескольких уровнях детализации с извлечением дополнительной информации из сигнала и повышением качества работы системы распознавания голоса диктора. Рассмотрим модель системы, где несколько уровней детализации исходного речевого сигнала используются для извлечения характеристических признаков говорящего.

Преимуществом данной модели по сравнению с предыдущей является повышение качества распознавания системы вследствие извлечения дополнительной информации из исходного речевого сигнала.

Недостатком является увеличение времени распознавания из-за дополнительных вычислений на нескольких уровнях детализации.

В работе [5] приведен пример реализации подобной системы, где в качестве характеристических признаков используются кепстальные коэффициенты, полученные из аппроксимаций исходного сигнала на разных уровнях детализации, и энтропия детализирующих коэффициентов вейвлет-преобразования. Авторы работы [5] выявили, что процент правильного распознавания по предлагаемому методу составляет 96,8 % в сравнении с 95,8 % для системы, построенной по классическому методу для сигнала без шума. При отношении сигнал–шум 20 дБ процент правильного распознавания составляет 91,6 %, в сравнении с системой, построенной по классическому методу (62,7 %) и по сравнению с системой с разделенными задачами фильтрации и распознавания (84,7 %).

Модель системы распознавания голоса диктора, использующая адаптивные деревья вейвлет-пакетов для извлечения характеристических признаков. Исследуя свойства речевых сигналов, выявили, что они имеют сложную структуру с быстро меняющимися характеристиками. Основным недостатком преобразования Фурье является отсутствие локализации по времени. БПФ с окном предполагает, что на анализируемом интервале сигнал стационарен, что не позволяет учитывать все особенности речевого сигнала.

Вейвлет-преобразование позволяет локализовать особенности речевого сигнала как по частоте, так и по времени и потенциально является более перспективным методом для решения задачи распознавания голоса диктора. Вместе с тем вейвлет-базисы значительно лучше приспособлены для фильтрации шума, что служит дополнительным аргументом для использования характеристических признаков, извлекаемых непосредственно из коэффициентов вейвлет-разложения.

Для решения задачи распознавания голоса диктора в зашумленной окружающей обстановке перспективной представляется система, в которой используют адаптивные деревья вейвлет-пакетов (рис. 3), которые имеют различное разрешение в разных частотных диапазонах. Например, согласно данным работы [6], диапазоны 100. . . 1000 Гц, 1000. . . 1500 Гц, 2000. . . 2500 Гц и 3000. . . 3500 Гц содержат больше характеристических признаков, чем диапазоны 1500. . . 2000 Гц, 2500. . . 3000 Гц и 3500. . . 4000 Гц и поэтому требуют более детального анализа.

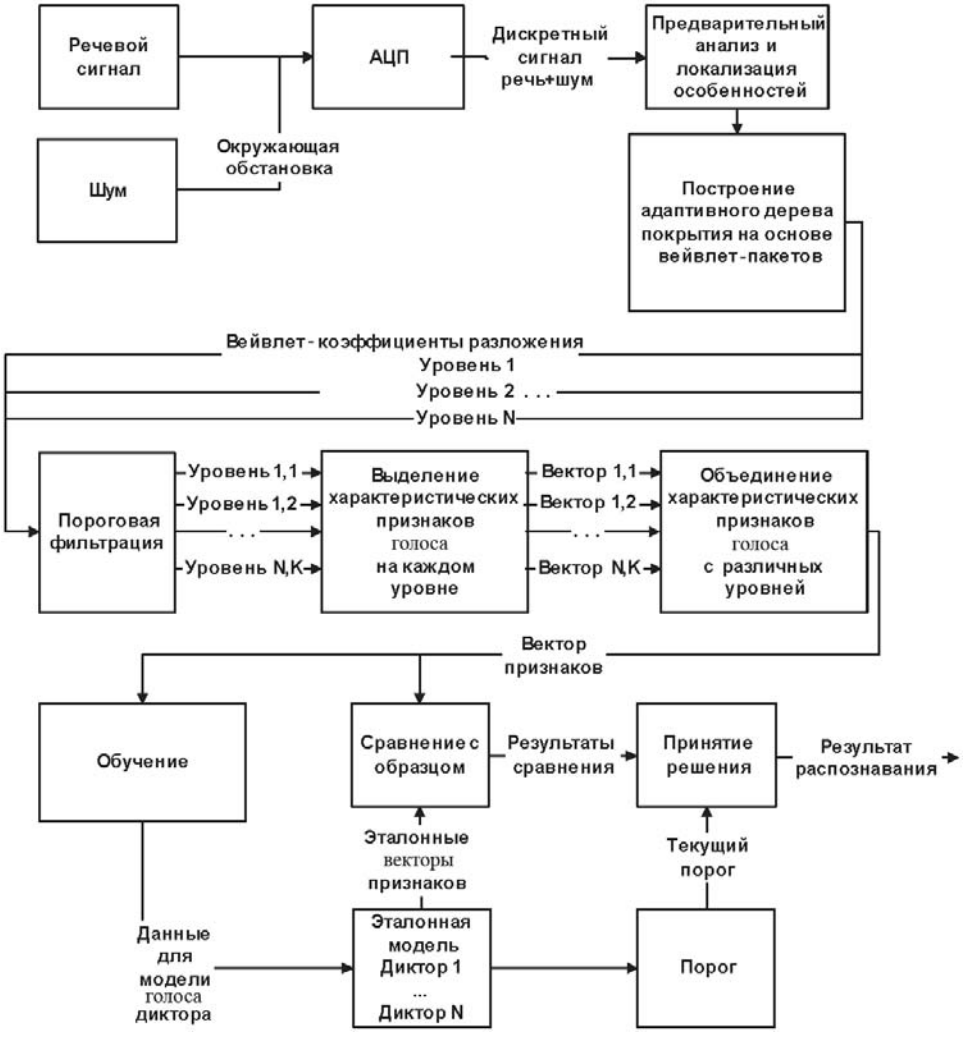


Рис. 3. Модель системы распознавания голоса диктора, использующая адаптивные деревья вейвлет-пакетов для извлечения характеристических признаков

Данная модель предполагает извлечение характеристических признаков непосредственно из вейвлет-коэффициентов на разных уровнях разложения, что позволит уменьшить объем дополнительных вычислений. С другой стороны, построение адаптивных деревьев с достаточным разрешением по частоте предполагает, что будет использовано как минимум 7 уровней разложения, что требует большого объема вычислительных ресурсов. В модели, представленной на рис. 4, достаточно использовать всего 3 уровня разложения.

Вычислительные затраты, которые требуются для достижения требуемого разрешения по частоте, являются, пожалуй, единственным недостатком представленной системы.

Выводы. Методы распознавания голоса диктора, основанные на преобразовании Фурье, хорошо справляются с поставленной задачей в идеализированной окружающей обстановке, однако в реальных систе-

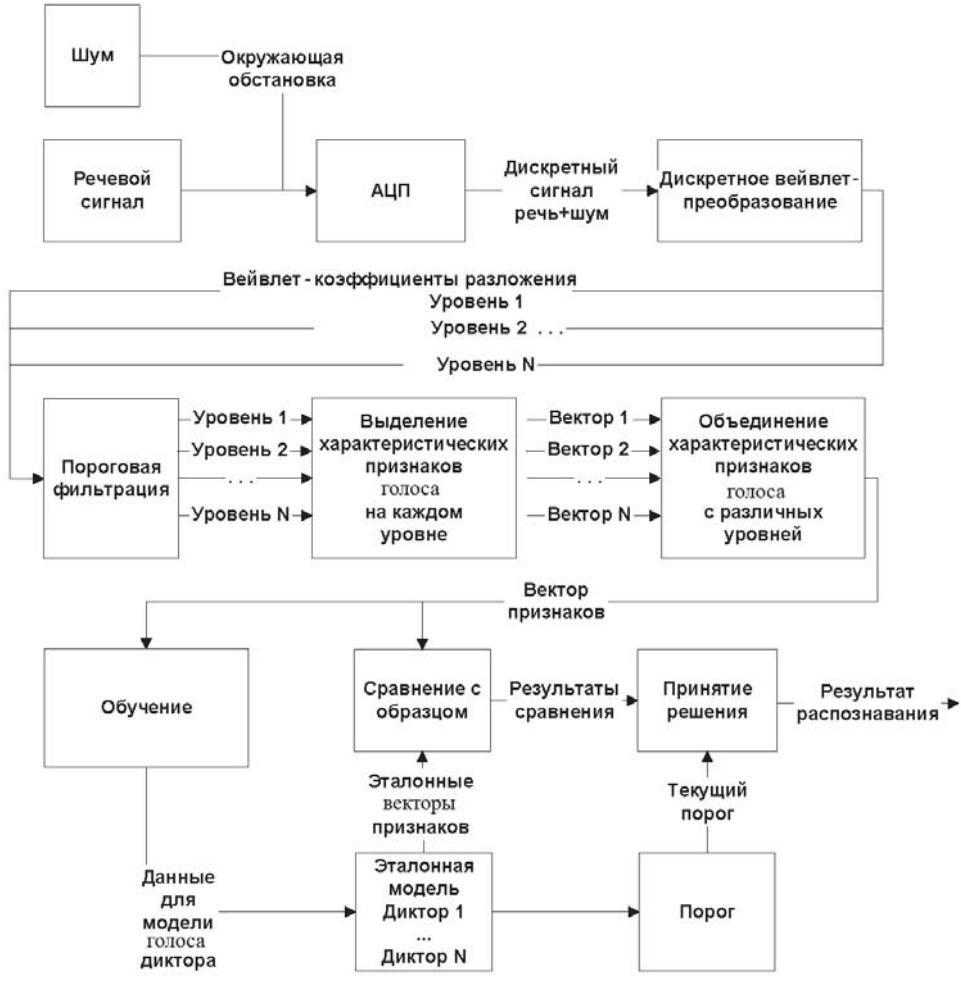


Рис. 4. Модель системы распознавания голоса диктора с совмещенными задачами фильтрации и распознавания

мах остро встает вопрос о снижении ошибок распознавания в зашумленной обстановке. Представленные методы позволяют существенно снизить влияние шума на качество распознавания голоса диктора и могут применяться как для доработки существующих систем, так и для построения принципиально новых систем, использующих самые современные и наиболее перспективные способы анализа и фильтрации речевых сигналов.

СПИСОК ЛИТЕРАТУРЫ

1. G a b o r D. Theory of communications / J. IEE, 93:429-457, 1946.
2. M a l l a C. Вейвлеты в обработке сигналов. – М.: Мир, 2005.
3. P r a v i n k u m a r P r e m a k a n t h a n & W a s f y . B . M i k h a e l . Speaker verification/identification and the importance selective feature extraction: Review. Department of Electrical Engineering, University of Central Florida, Orlando.

4. Farooq O. and Datta S. A novel wavelet based pre-processing for robust features in ASR. Department of Electronic and Electrical Engineering Loughborough University, Loughborough, LE11 3TU, UK.
5. Ching-Tang Hsieh, Eugene Lai and You-Chuang Wang. Robust speaker identification system based on wavelet transform and Gaussian mixture model // Journal of Information Science and Engineering 19, 267–282 (2003).
6. M. Siafarikas, T. Ganchev, N. Fakotakis. Objective wavelet packet features for speaker verification. Wire Communications Laboratory, University of Patras, Rio-Patras 26500, Greece.

Статья поступила в редакцию 5.11.2007

Владимир Борисович Кропотов родился в 1980 г., окончил Рязанскую государственную радиотехническую академию в 2003 г. Аспирант кафедры “Информационная безопасность” МГТУ им. Н.Э. Баумана. Автор пяти научных работ в области информационной безопасности.

V.B. Kropotov (b. 1980) graduated from the Ryazan State Radio Engineering Academy in 2003. Post-graduate of “Information Security” department of the Bauman Moscow State Technical University. Author of 5 publications in the field of information security.

Николай Викторович Медведев родился в 1954 г., окончил в 1977 г. МВТУ им. Н.Э. Баумана. Канд. техн. наук, зав. кафедрой “Информационная безопасность” МГТУ им. Н.Э. Баумана. Автор около 50 научных работ в области исследования и разработки защищенных систем автоматической обработки информации.

N.V. Medvedev (b. 1954) graduated from the Bauman Moscow Higher Technical School in 1977. Ph. D. (Eng.), head of “Information Security” department of the Bauman Moscow State Technical University. Author of about 50 publications in the field of study and development of secured systems of automatic data processing.

Игорь Иванович Троицкий родился в 1955 г., окончил в 1978 г. Московский инженерно-физический институт. Канд. техн. наук, доцент кафедры “Информационная безопасность” МГТУ им. Н.Э. Баумана. Автор около 20 научных работ в области информационной безопасности и исследования систем обработки информации и управления.

I.I. Troitskii (b. 1955) graduated from the Moscow Engineering and Physical Institute in 1978. Ph. D. (Eng.), assoc. professor of “Information Security” department of the Bauman Moscow State Technical University. Author of about 20 publications in the field of information security and study of systems of data processing and management.