БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УДК 543.426

№ госрегистрации 20002165

УТВЕРЖДАЮ

Проректор по научной работе,
д-р хим. наук

_____С.К. Рахманов

« » декабря 2000 г.

ОТЧЕТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НЕЙРОСЕТЕВЫЕ МОДЕЛИ И АЛГОРИТМЫ АНАЛИЗА ХАРАКТЕРИСТИК ПОЛУПРОВОДНИКОВЫХ ПРИБОРОВ

(заключительный)

г.б. НИР №682/18

Зав. кафедрой
системного анализа
д-р физ.-мат. наук, профессор В.В. Апанасович

Научный руководитель
мл. научн. сотр., аспирант О.К. Барановский

Руководитель (куратор) проекта
старший преподаватель В.М.Лутковский

Минск 2000

СПИСОК ИСПОЛНИТЕЛЕЙ

Руководитель проекта	В.М. Лутковский (Введение, заключение)
Научный руководитель работы,	О.К.Барановский (1 , 3.1)
Ответственный исполнитель, Мл. науч. сотр.	П.В.Назаров (2.1,2.2, 3.2, 4)
Инженер	А.Ю. Балахонцев (2.3, 3.3)
Нормоконтролер	Т.Н. Долгая

РЕФЕРАТ

Отчет 74 страницы, 30 рисунков, 65 формул, 38 источников.

ПОЛУПРОВОДНИКОВЫЕ ПРИБОРЫ, НЕЙРОННЫЕ СЕТИ, РАДИАЛЬНО-БАЗИСНЫЕ ФУКЦИИ, МОДЕЛИРОВАНИЕ, ВОССТАНОВЛЕНИЕ ПАРАМЕТРОВ, АППРОКСИМАЦИЯ, ПРОГНОЗИРОВАНИЕ, КОРРЕКЦИЯ СИГНАЛОВ.

Объектом исследований являются полупроводниковые приборы с лавинным умножением зарядов.

Основная цель предлагаемого исследования состоит в разработке и исследовании нейросетевых моделей полупроводниковых приборов. Построение таких моделей позволяет осуществлять коррекцию и прогнозирование выходных сигналов, а также производить восстановление параметров прибора по известным выходным характеристикам.

Основные методы базируются на применении теории нейронных сетей, линейной алгебры и вычислительной математики.

Построены модели на основе нейронных сетей для аппроксимации и прогнозирования выходных сигналов полупроводниковых диодов. Проведен анализ эффективности этих моделей на основе машинного эксперимента с использованием данных физического эксперимента.

Построена нейронная сеть для определения режимов работы и характеристик полупроводниковых диодов. Данная сеть протестирована на основе экспериментальных данных.

Разработаны нейросетевые модели для коррекции сигналов в системах обработки информации с целью устранения аппаратных искажений. Построена модель многоэлементного фотоприемника с возможностью устранения систематических ошибок. Разработана модель для коррекции инерционности фотодетекторов. Предложенные модели использованы в коррекции данных физического эксперимента и показали хорошие результаты.

СОДЕРЖАНИЕ

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ, УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ, ЕДИНИЦ И ТЕРМИНОВ

ИНС – искусственная нейронная сеть;

ЛФД – лавинный фотодиод;

МСП – многослойный персептрон;

МЭФП – многоэлементный фотоприёмник;

ОПЗ – область пространственного заряда;

ПЗС – прибор с зарядовой связью;

РБФ – радиально-базисная функция;

trace – след матрицы;

diag – сумма диагональных элементов.

ВВЕДЕНИЕ

Создание моделей приборов и систем является неотъемлемым этапом современных технологий проектирования и производства. Моделирование позволяет значительно сократить материальные затраты на всех этапах проектирования новой техники, так как оно предполагает замену реального объекта некоторой другой (во многих случаях нематериальной) системой, повторяющей существенные свойства оригинала. Однако, усложнение структуры полупроводниковых приборов и интегральных схем требует создания более эффективных технологий построения моделей [1-3].

Перспективным направлением разработки таких технологий представляется использование новых подходов, в основе которых лежат теория и практическое применение искусственных нейронных сетей [4-5]. В рамках настоящей работы предлагается два различных подхода к разработке нейросетевых алгоритмов.

Первый подход основывается на накопленном опыте математического моделирования и направлен на представление известных математических моделей в нейросетевом базисе, позволяющем распараллелить алгоритм вычислений. Необходимо отметить, что такой подход не отрицает традиционных методов математического и физического моделирования, направленных на изучение внутренней структуры и физических процессов внутри исследуемого объекта, и способствует ускорению получения результатов вычислительного эксперимента [6-9].

Второй подход не требует детального описания внутренней структуры и параметров моделируемого полупроводникового прибора, но предполагает наличие данных о взаимосвязи наблюдаемых характеристик полупроводниковых приборов. В этом смысле его можно назвать непараметрическим [10]. Такой подход наиболее эффективен для построения моделей новых приборов, так как применим при наличии достаточно большого объема экспериментальных данных и отсутствии строгой теории и разработанных средств проектирования. Такой подход основан на таких уникальных свойствах нейронных сетей как способность к обучению и обобщению [4-5].

Первый раздел работы посвящён анализу современного состояния методов математического моделирования процессов в полупроводниковых приборах.

В следующих двух разделах исследованы особенности архитектур нейронных сетей, а также алгоритмы моделирования приборов, построенные на их основе.

В третьем и четвертом разделах рассмотрены примеры практического применения разработанных нейросетевых алгоритмов. Их выбор обусловлен доступностью данных о характеристиках моделируемых приборов. В качестве таких примеров в работе приведены:

- результаты оценивания параметров, условий и режимов работы кремниевых диодов по сигналам, генерируемым в режиме микроплазменного пробоя;
- результаты прогнозирования динамики изменения режимов работы исследованных полупроводниковых приборов;
- методы коррекции сигналов приборов, используемых в качестве детекторов.

В заключении подведены итоги и перечислены наиболее важные результаты работы.

1. МОДЕЛИ И АЛГОРИТМЫ АНАЛИЗА ХАРАКТЕРИСТИК ПОЛУПРОВОДНИКОВЫХ ПРИБОРОВ

1.1. Введение

В данном разделе анализируются классические методы моделирования, используемые в процессе исследования структуры и принципов работы полупроводниковых приборов и электронных систем. В математическом смысле выбор модели – это выбор соответствующего уравнения, поэтому значительное внимание уделяется вопросам математического описания процессов в полупроводниковых приборах. Высокая сложность математических моделей полупроводниковых приборов объясняется многообразием процессов, протекающих в их объеме и на поверхности. Например, процессы в обратносмещенном электронно-дырочном переходе с учетом микроплазм описывается системой уравнений, состоящей из уравнения баланса для средней по толщине области пространственного заряда (ОПЗ) концентрации электронов, уравнения непрерывности полного тока в квазинейтральных n- и p-областях структуры и упрощенного усредненного по толщине ОПЗ уравнения теплопроводности.

Наиболее важной чертой традиционного подхода к построению моделей новых полупроводниковых приборов является стремление описать их поведение с помощью одной или нескольких известных математических моделей. Это вполне логично и эффективно с точки зрения снижения затрат на этапе создания модели нового прибора. Однако, структура таких моделей усложняется для новых поколений приборов и на этапе их применения часто возникает вопрос о замене таких моделей с целью снижения вычислительных затрат.

В подавляющем большинстве случаев при проектировании интегральных схем с высокой степенью интеграции и сложных электронных систем применяются численные методы. Поэтому значительное место в данном разделе уделяется методам, основанным на применении ЭВМ. Рассматриваются также процедуры построения моделей, используемых для прогнозирования поведения приборов во времени и построения закономерностей их функционирования. При этом исследуется возможность использования информации о вторичных процессах в полупроводниковых приборах (генерации микроплазм и избыточного шума) для оценки состояния и прогнозирования поведения этих приборов.

1.2. Методы моделирования при анализе полупроводниковых приборов

В зависимости от характера изучаемых процессов в приборе или системе все методы моделирования могут быть разделены на детерминированные и стохастические, статические и динамические, дискретные, непрерывные и дискретно-непрерывные. Детерминированное моделирование отображает детерминированные процессы, то есть процессы, в которых предполагается отсутствие всяких случайных воздействий; стохастическое моделирование отображает вероятностные процессы и события. В этом случае анализируется ряд реализаций случайного процесса и оцениваются средние характеристики, то есть набор однородных реализаций. Статическое моделирование служит для описания поведения объекта в какой-либо момент времени, а динамическое моделирование отражает поведение объекта во времени. Дискретное моделирование служит для описания процессов, которые предполагаются дискретными, соответственно непрерывное моделирование позволяет отразить непрерывные процессы в системах, а дискретно-непрерывное моделирование используется для случаев, когда хотят выделить наличие как дискретных, так и непрерывных процессов [11].

В зависимости от формы представления объекта (системы) можно выделить мысленное и реальное моделирование. Мысленное моделирование может быть реализовано в виде наглядного, символического и математического. Остановимся на математическом моделировании.

В свою очередь, математическое моделирование для исследования характеристик процесса функционирования систем можно разделить на аналитическое, имитационное и комбинированное.

Для аналитического моделирования характерно то, что процессы функционирования элементов системы записываются в виде некоторых функциональных соотношений (алгебраических, интегрально-дифференциальных, конечно-разностных и т.п.) или логических условий. Аналитическая модель может быть исследована следующими методами:

а) аналитическим, когда стремятся получить в общем виде явные зависимости для искомых характеристик;

б) численным, когда не имея общих решений уравнений, стремятся получить числовые результаты при конкретных начальных данных;

в) качественным, когда не имея решения в явном виде, можно найти некоторые свойства решения (например, оценить устойчивость решения).

Наиболее полное исследование процесса функционирования системы можно провести, если известны явные зависимости, связывающие искомые характеристики с начальными условиями, параметрами и переменными системы. Однако, такие зависимости удается получить только для сравнительно простых случаев. При усложнении систем исследование их аналитическим методом наталкивается на значительные трудности, которые часто бывают непреодолимыми. Поэтому, желая использовать аналитический метод, в этом случае идут на существенное упрощение первоначальной модели, чтобы иметь возможность изучить хотя бы общие свойства системы. Численный метод, эффективно используемый с применением ЭВМ, позволяет исследовать по сравнению с аналитическим методом более широкий класс систем, но при этом полученные решения носят частный характер.

Высокая сложность математических моделей полупроводниковых приборов объясняется многообразием процессов, протекающих в их объеме и на поверхности.

Примером тому служит полупроводниковый диод в режиме обратного смещения. В большинстве практических применений применяется модель идеализированного *p*-*n*-перехода, не учитывающая его технологические особенности и ряд важных физических явлений, определяющих его рабочие характеристики [1-2]. Однако известно, что при лавинном пробое обратно смещенных *p*-*n*-переходов наблюдается образование локализованных областей высокой плотности тока, которые получили названия микроплазм [12]. При этом, теоретические модели [13] не позволяли описывать спонтанное включение-выключение микроплазм.

В работе [14] предложена теоретическая модель, объясняющая спонтанное включение-выключение саморазогревом *p*-*n*-структуры. Образование микроплазм описывается системой уравнений (1), состоящей из уравнения баланса для средней по толщине области пространственного заряда (ОПЗ) концентрации электронов, уравнения непрерывности полного тока в квазинейтральных *n*- и *p*-областях структуры и упрощенного усредненного по толщине ОПЗ уравнения теплопроводности:

$$\begin{aligned}
\frac{\partial n}{\partial t} &= D\Delta_{\perp} n + n\nu_i\left(n, V_i\right) - \frac{n}{\tau_n} + G_T, \\
C\frac{\partial V_i}{\partial t} &= \sigma\tilde{W}\Delta_{\perp} V_i - j + (V - V_i)\rho^{-1}, \\
\tau_T\frac{\partial T}{\partial t} &= \lambda^2\Delta_{\perp} T - \left(T - T_i\right) + \tilde{C}V_i n,
\end{aligned} \tag{1}$$

где ν_i – средняя по толщине ОПЗ скорость ионизации носителей заряда, V_i – падение напряжения на ОПЗ p–n-перехода, $j = en v_n$ – плотность лавинного тока, D и v_n – коэффициент диффузии и дрейфовая скорость электронов в ОПЗ, $\tau_n = \varpi / v_n$ – время пролета носителей через ОПЗ, ϖ – толщина ОПЗ, C – удельная емкость *p–n*-структуры, $\Delta_\perp \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$, ось z выбрана по нормали к плоскости *p–n*-перехода, $\tilde{W}$ – эффективная толщина растекания тока в квазинейтральных p- и n-областях, σ – проводимость квазинейтральной области, $\rho = \tilde{W}/\sigma$, V – полное падение напряжения на p–n-структуре, G_T – скорость тепловой и туннельной генерации в ОПЗ, τ_T, λ - время и длина тепловой релаксации, $\tilde{C} = e v_n R_T$ – удельное тепловое сопротивление образца, T_i – температура эффективного термостата (подложки).

Система уравнений (1) не решается аналитическими методами, но ее легко свести к системе разностных уравнений, удобной для численного решения.

Теоретическая модель, рассмотренная выше, имеет большое количество параметров и предполагает постоянный профиль неоднородностей, на которых образуются микроплазмы. Однако, воздействие таких внешних факторов, как жесткое ионизирующее излучение, может вызывать увеличение количества неоднородностей p-n-структуры, перераспределяя участки образования микроплазм [15]. Поэтому необходим учет вероятностного изменения характера образования лавин в микроплазмах [16], который реализуется при имитационном моделировании.

При имитационном моделировании алгоритм, реализующий модель, воспроизводит процесс функционирования системы во времени. При этом программно моделируются элементарные явления, составляющие процесс, с сохранением их логической структуры и последовательности протекания во времени. Имитационные модели позволяют достаточно просто учитывать такие факторы, как наличие дискретных и непрерывных элементов, нелинейные характеристики элементов системы, многочисленные случайные воздействия и другие, которые часто создают трудности при аналитических исследованиях.

Когда результаты, полученные при воспроизведении на имитационной модели процесса функционирования системы, являются реализациями случайных величин и функций, тогда для нахождения характеристик процесса требуется его многократное воспроизведение с последующей статистической обработкой информации и

целесообразно в качестве метода машинной реализации имитационной модели использовать метод статистического моделирования или метод Монте-Карло.

Одним из основных преимуществ метода Монте-Карло является то, что он позволяет проводить моделирование различных физических процессов на микроскопическом уровне. Кинетическое моделирование процессов переноса носителей заряда в полупроводниках методом Монте-Карло дает возможность непосредственно рассчитывать электрофизические параметры полупроводниковых материалов и влияние на них различных факторов, в частности, сильных электрических полей. Остановимся на основных особенностях простейшего базового алгоритма моделирования эти методом переноса носителей заряда в полупроводниках. Основная суть рассматриваемого метода состоит в прослеживании движения электрона в пространстве волновых векторов $\bar{k}\left(k_x, k_y, k_z\right)$ и координатном пространстве $\bar{r}(x, y, z)$. При этом свободное движение электрона в электрическом поле прерывается случайным образом процессами его рассеяния фононами, примесями, неоднородностями поверхности и так далее. В момент рассеяния электрон мгновенно переходит из одной точки волнового пространства в другую, в то время как его положение в пространстве координат остается прежним [3].

Примером такой задачи может служить разработка лавинных фотодиодов с импульсной характеристикой минимальной длительности для волоконно- оптических систем связи [17].

Имитационная модель, используемая для изучения зависимости характеристик рассмотренного прибора от параметров структуры, в своем ядре содержит моделирование траектории носителей заряда с учетом воздействия электрических полей, фононов и так далее. Такая модель описывает системы с $N \approx 10^{22}$ частиц, в которой необходимо использовать метод Монте-Карло для имитации процессов диффузии, движения под воздействием электрического поля, актов соударения с атомами решетки и дефектами более чем 10^{17}-10^{18} носителей заряда. Так как такой подход требует огромнейших размеров операционной памяти и машинного времени, то необходимо будет использовать системы меньших размеров, для которых возникает нетривиальный вопрос об экстраполяции на большие системы [18]. Процесс моделирования при этом включает определение траектории одного или нескольких носителей, средней энергии, тока, средней дрейфовой скорости и так далее. Другими

словами, метод Монте-Карло применим в тех случаях, когда изучены физические процессы, определяющие работу данного прибора.

Строго говоря, процессы в твердотельных структурах протекают в трехмерном физическом пространстве и во времени. Однако, требование огромных вычислительных ресурсов в задачах трехмерного моделирования вызывает практический интерес к привлечению более мощных вычислительных средств, а при их отсутствии - к использованию двумерного приближения, при котором процессы рассматриваются в сечении микроэлектронного фрагмента [1,2].

1.3. Особенности анализа нестационарных процессов в приборах

1.3.1. Моделирование эволюционных процессов

В каждой системе, проявляющей свойства “роста”, необходимо вводить соответствующие переменные в качестве показателей этого роста. Эти показатели мы должны рассматривать как координаты состояния рассматриваемой системы, так как в их значениях, очевидно, аккумулируются определенные влияния. Тем самым конкретную реализацию процесса роста можно описывать как траекторию $z(t)$ в постоянном пространстве состояний [19].

Особенно важными эволюционными процессами являются процессы роста, характеризуемые свойствами монотонности их траекторий. Это могут быть, например, размножение носителей заряда в лавине или рост дефектов в процессе облучения образца потоками жесткого излучения.

Эволюционные процессы можно описывать только в рамках сложных систем, когда возможно представить такую систему в виде определенной структуры графа, причем элементарные подсистемы являются вершинами этого графа. Элементарные характеристики эволюции описываются посредством дифференциальных уравнений следующего вида:

$$\frac{dx}{dt} = f(x, y_+, y_-). \qquad (2)$$

Здесь y_+ обозначает влияние, стимулирующее рост, а у- - влияние, тормозящее рост. Когда y_+, y_- жестко заданы заранее, мы приходим к автономному поведению системы, которая описывается дифференциальными уравнениями вида:

$$\frac{dx}{dt} = f(x_+, x_-). \tag{3}$$

Неравенство $f \geq 0$ является необходимым и достаточным условием для растущей системы. Пусть $f \geq 0$, $g \geq 0$ и правая часть выражения (2) имеет разделение переменных, тогда

$$\frac{dx}{dt} = f(x)g(y_+, y_-). \tag{4}$$

Если $g(y_+, y_-)$ интегрируема, то решение дифференциального уравнения имеет вид

$$x(t) = F^{-1}\left(F(x_0) + \int_0^t g(y_+, y_-)dt \right). \tag{5}$$

В выражении (5) $F(x) = \int_0^x du / f(u)$. Характер поведения $x(t)$ зависит от значений функций $F(x_0)$ и второго члена правой части уравнения (5). При допущении $y_- = 0$ следует, что $F(x)$ монотонно возрастает и, следовательно, $x(t)$ также будет возрастать. Гораздо более сложное поведение $x(t)$ будет наблюдаться при $y_- \neq 0$. $x(t)$ будет определяться соотношением y_+, y_-.

В реальных системах, например, когда через $x(t)$ обозначается количество носителей заряда в лавине при микроплазменном пробое p-n-перехода, разделение переменных в правой части выражения (2) может быть затруднено. Это происходит вследствие зависимости y_+, y_- от числа носителей заряда в лавине.

1.3.2. Модельные методы нечеткой идентификации

Как бы нам не хотелось применять детерминистские модели, реальные явления всегда представляют собой смешение детерминированных и случайных явлений. Поэтому, для каждого конкретного реального объекта необходимо настолько определить поведение модели, чтобы оставалось только уточнить значения еще

нескольких свободных параметров с помощью результатов экспериментов над моделью [19].

Продемонстрируем такой образ действий на примере идентификации временного тренда. При этом преследуется цель предсказать будущее поведение временного ряда.

Примерный вид детерминированной модели будет иметь вид

$$x_M(t) = f(t, p) \tag{6}$$

с набором $p = p_1, p_2, ..., p_k$ параметров, еще подлежащих определению. Предполагая, что временной ряд является показателем автономного процесса роста, можно назначить детерминированную функцию тренда.

Какого рода информацию можно ожидать от измерений в общем случае? В множестве интервалов времени t_i заключена информация в виде интервалов нечеткости. Следовательно, мы знаем, что истинные выборочные значения $x(t_i)$ временного ряда принадлежат этим интервалам, причем относительно их положения внутри этих интервалов могут иметься различные неточности. Различные возможности для точек интервала можно описать с помощью функций распределения, однако в нашем случае применим концепцию нечеткости Заде [20]. Интервальная информация осмысливается с помощью распределения нечеткости $\rho_\Delta(u_i, x(t_i))$. Характер этого распределения должен в агрегированной форме хранить априорную информацию о временном ряде к моменту времени $t = t_i$. Элементарные распределения нечеткости u_i связаны с эталонным значением детерминированной модели с помощью шумовой переменной n_i:

$$u_i = n_i + f(t_i, p). \tag{7}$$

Предположим, что нечеткое приращение n не зависит от времени. Тогда оценка для получения качественной информации о распределении нечеткости ошибки нечеткой модели представляет собой нечеткую дизъюнкцию в качестве нечеткого аналога теоретико-множественного объединения:

$$\rho^*(n;p) = disj \cdot \rho_\Delta(u_i - f(t_i, p)). \quad (8)$$

Для агрегированной функции $\rho^*(n;p)$ необходимо определить центральную точку (медиану) и меру ее ширины. Для предсказания значения ряда в последующие моменты времени t' рассчитываем $f(t', p)$, используя полученное значение как центральную точку для предсказания. Затем используем ширину агрегированного распределения ошибки, предсказываем величину доверительного интервала вокруг центрального значения.

В общем случае нечеткие приращения n_i могут зависеть от времени. Тогда необходимо оценивать по крайней мере двумерные нечеткие распределения на основе двумерных распределений нечеткости $\rho_\Delta(u_i - f(t_i, p), u_j - f(t_j, p))$. Тем самым возможно также зависящее от времени предсказание нечеткости.

Еще более сложные зависимости получаются при условии, что предсказываемый процесс имеет вероятностную функцию описания. Примером тому служит последовательность микроплазменных импульсов в генераторных диодах.

1.4. Кибернетическое моделирование

Описанные выше проблемы указывают на необходимость разработки более эффективных способов моделирования. Значительный интерес в этом отношении представляет кибернетическое моделирование. В этом случае реальный объект рассматривают как “черный ящик”, имеющий ряд входов и выходов. При этом снимается требование подобия реальных физических процессов и процессов, происходящих в моделях, и моделируются лишь функциональные связи между выходами и входами.

Чаще всего при использовании кибернетических моделей проводят анализ поведенческой стороны объекта при различных воздействиях внешней среды. Таким образом, в основе кибернетических моделей лежит отражение некоторых информационных процессов управления, что позволяет оценить поведение реального объекта. Для построения имитационной модели в этом случае необходимо выделить исследуемую функцию реального объекта, попытаться формализовать эту функцию в виде некоторых операторов связи между входом и выходом и воспроизвести на имитационной модели данную функцию, причем на базе совершенно иных математических соотношений и, естественно, иной физической реализации процесса.

Характеризуя процесс кибернетического моделирования обращают внимание на следующие обстоятельства. Модель, будучи аналогом исследуемого явления, никогда не может достигнуть степени сложности последнего. При построении модели прибегают к известным упрощениям, цель которых - стремление отобразить не весь объект, а с максимальной полнотой охарактеризовать некоторый его “срез”. Задача заключается в том, чтобы путем введения ряда упрощающих допущений выделить важные для исследования свойства. Создавая кибернетические модели, выделяют информационные свойства сигналов или систем. Все иные стороны этого объекта остаются вне рассмотрения [21].

Преимущества кибернетического моделирования реализуются при использовании непараметрических алгоритмов нейросетевого моделирования, рассматриваемых в следующем разделе.

1.5. Выводы

В данном разделе рассмотрены традиционные методы моделирования, используемые для анализа процессов в полупроводниковых приборах, и выявлены ограничения, связанные с использованием таких методов. Основным ограничением является отсутствие полной информация о протекающих в приборах процессах, а также отсутствие в ряде случаев информации о структуре и параметрах самих приборов. Это подтверждается как использованием аналитических, так и численных математических методов моделирования. Ситуация может быть разрешена путем применения имитационных моделей с использованием методов Монте-Карло. Однако такие модели требуют недостижимых вычислительных ресурсов и быстродействия при современном уровне развития ЭВМ. Кроме того, каждый из этапов реализации таких моделей в виде компьютерных программ, проведения вычислительных экспериментов с их использованием, а также обработки полученных данных, требует привлечения специалистов высокой квалификации и связан с высокими временными и интеллектуальными затратами.

Требование огромных вычислительных ресурсов в задачах трехмерного моделирования вызывает практический интерес к привлечению более мощных вычислительных средств, а при их отсутствии - к использованию упрощенных моделей.

2. НЕЙРОСЕТЕВОЙ ПОДХОД К МОДЕЛИРОВАНИЮ

2.1. Введение

В настоящее время область искусственные нейронных сетей (ИНС) является одной из наиболее бурно развивающихся областей науки. Появившись в конце 40-х на стыке кибернетики, математики и биологии, она переживала взлёты и падения, однако, благодаря работам энтузиастов, доказавших её преимущества, с середины 80-х годов наблюдается рост интереса к этой области.

Особенность нейросетевого подхода к моделированию заключается в использовании универсальных программных или аппаратных средств, настраиваемых под конкретную задачу путем их обучения по специально подготовленному множеству. В рамках данного раздела предполагается, что ИНС моделируются на стандартном персональном компьютере, однако в равной степени они могут быть реализованы на специальных микросхемах, нейроплатах или нейрокомпьютерах [4].

Целесообразность перехода к нейросетевым алгоритмам моделирования может быть обусловлена следующими причинами:

- необходимо существенно ускорить процесс проведения вычислительного эксперимента;
- необходимо снизить затраты на этапе создания программной реализации моделей;
- не удается привести исходную задачу к описанию последовательности операций над входными данными или другому формализованному виду, позволяющему построить алгоритм ее решения на ЭВМ.

Другими словами, искусственные нейронные сети позволяют решать многие трудно формализуемые задачи и отличаются достаточно низкими временными затратами на полный цикл их решения. В процессе построения модели ИНС приводит исходную задачу к специальному нейросетевому базису $\left\{\sum \bar{a}\,\bar{x}\right\}$ с последующей классификацией входных данных по указанному базисному разложению. Следует заметить, что многие уравнения математической физики могут быть представлены в таком базисе [4]. Именно это может быть использовано в качестве отправного момента при разработке алгоритмов моделирования полупроводниковых приборов.

Выбор универсальных программных средств, имитирующих ИНС определяется классом решаемых задач, поэтому в данном разделе рассмотрены потенциальные возможности многослойных персептронов и нейронных сетей, построенных на основе

радиально-базисных функций. Под точностью моделирования понимается точность аппроксимации тех или иных характеристик приборов.

Рассмотрим основные принципы построения ИНС.

2.2. Основы теории нейронных сетей

Искусственная нейронная сеть представляет собой совокупность процессорных элементов (нейронов). Искусственный нейрон состоит из набора межсоединений, сумматора и нелинейного оператора. Значение каждого входа X_i умножается на свой вес W_i, затем суммируется и из суммы вычитается порог θ. После этого на результат действует некоторая активационная функция F и преобразованная таким образом информация подаётся на выход нейрона Y, как показано на рисунке1 [5].

Таким образом, выход каждого нейрона вычисляется по формуле:

$$Y = F\left\{\sum_{i=1}^{n}\left(X_i W_i\right) - \theta\right\}. \qquad (9)$$

В дальнейшем условимся обозначать сумму произведений $X_i W_i$ через переменную NET.

В данной работе использованы два класса ИНС: многослойные персептроны (МСП) и сети с радиальными базисными функциями (РБФ-сети). Рассмотрим структуры этих сетей и способы их применения.

2.3. Многослойные персептроны

2.3.1. Структура сети

Многослойным персептроном называют нейронную сеть, изображённую на рисунке 2, нейроны в которой расположены слоями, причём нейроны каждого слоя связаны только с нейронами предыдущего слоя.

Модель искусственного нейрона

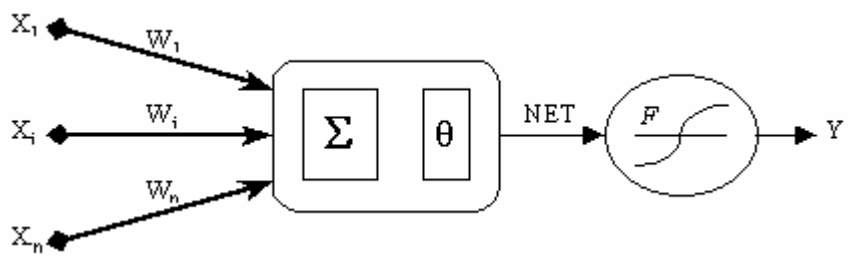


Рис. 1

Многослойный персептрон

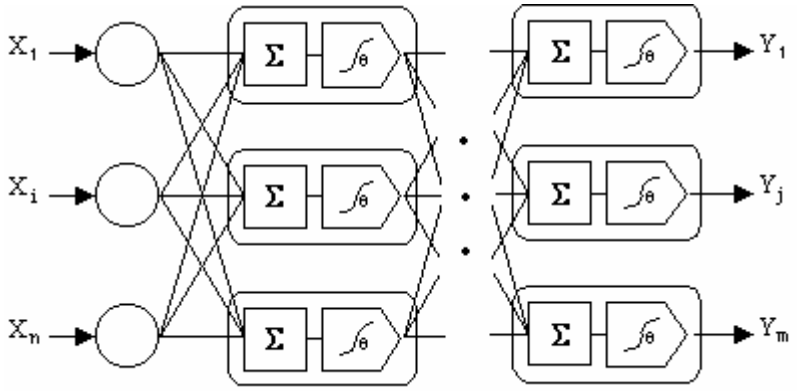


Рис. 2

Для сетей этого класса существует довольно хороший "классический" метод обучения – так называемый метод обратного распространения ошибки (*Back Propagation Error*), который относится к разряду методов градиентного спуска [5,22,23]. Он очень важную роль в пробуждении интереса к нейронным сетям в начале 80-х, а его модификации и по сей день остаются лучшими для рассматриваемого класса сетей.

2.3.2. Обучение сети

С использованием процедуры обратного распространения решаются задачи, в которых для некоторого входного вектора $X(x_1, \ldots, x_n)$ сеть на выходе формирует требуемый вектор $Y(y_1, \ldots, y_n)$. На этапе обучения используется ряд пар векторов (X, Y^*), называемых обучающей парой, где Y^* – целевой вектор.

Используя обучающие пары, сеть подстраивает свои веса таким образом, чтобы адекватно реагировать на входной вектор X. Рассмотрим работу алгоритма обратного распространения ошибки в случае, когда в качестве активационной функции выступает сигмоидальная функция.

При инициализации сети случайным образом задаются начальные веса и сдвиги сети. Алгоритм обучения сети обратного распространения включает в себя следующие шаги:

1. Выбрать очередную обучающую пару (X,Y^*) из обучающего множества и подать входной вектор X на вход сети.
2. Вычислить выход сети Y.
3. Вычислить разность между реальным (вычисленным) выходом сети и требуемым выходом (целевым вектором обучающей пары).
4. Подкорректировать веса сети так, чтобы минимизировать ошибку.

Повторить шаги с 1 по 4 для каждого вектора обучающего множества до тех пор, пока ошибка на всем множестве не достигнет приемлемой величины.

Шаги 1 и 2 используются как на этапе обучения сети, так и при функционировании уже обученной сети.

Вычисления в сети выполняются послойно. На шаге 3 каждый из выходов сети Y вычитается из соответствующей компоненты целевого вектора с целью получения ошибки. Эта ошибка используется на шаге 4 для коррекции весов сети, причем величина изменений определяются алгоритмом обучения.

Шаги 1 и 2 можно рассматривать как "проход вперед", а 3 и 4 — как проход назад, так как сигнал ошибки распространяется обратно по сети и используется для подстройки весов. Эти два прохода можно выразить математически.

На входе имеем вектор *X*, на основе которого вычисляется выходной вектор *Y*. Вектор *Y* вычитается из целевого вектора Y^* с целью получения ошибки ε:

$$\varepsilon = Y^* - Y\,. \tag{10}$$

Величина *NET* каждого нейрона первого слоя вычисляется как взвешенная сумма входов нейрона. Затем активационная функция *F* сжимает *NET* и дает величину *OUT* для каждого нейрона в этом слое.

Полученное выходное множество *OUT* является входом для следующего слоя. Процесс повторяется слой за слоем, пока не будет получено заключительное множество сети.

На этапе обратного прохода происходит подстройка весов выходного слоя. Так как для каждого нейрона выходного слоя задано целевое значение, то подстройка весов легко осуществляется с помощью дельта правила. Внутренние слои не имеют целевых значений и называются скрытыми слоями.

Процесс подстройки одного веса от нейрона *p* в скрытом слое *j* к нейрону *q* в выходном слое строится сдедующим образом. Выход *OUT* слоя *k* вычитаясь из целевого значения Y^*, дает ошибку, которая умножается на производную сжимающей функции (в нашем случае $OUT(1\text{-}OUT)$), вычисленную для этого нейрона слоя *k*, давая, таким образом, величину

$$\delta = OUT(1-OUT)(Y^* - OUT)\,. \tag{11}$$

Затем δ умножается на величину *OUT* нейрона *j*, из которого выходит рассматриваемый вес. Это произведение в свою очередь умножается на коэффициент обучения η (0,01≤η≤1) и результат прибавляется к весу.

$$\Delta w_{pq,k} = \eta\delta_{q,k} \cdot OUT_{p,q}\,, \tag{12}$$

где $\delta_{q,k}$ - величина δ для нейрона q в выходном слое k; $OUT_{p,q}$ - величина выхода для нейрона в скрытом слое j.

$$w_{pq,k}^{(n+1)} = w_{pq,k}^{n} + \Delta w_{pq,k}, \tag{13}$$

где $w_{pq,k}^{n}$ – величина веса от нейрона в скрытом слое k к нейрону q в выходном слое на шаге n, $w_{pq,k}^{(n+1)}$ – величина веса на шаге $n+1$ после коррекции. Такая же процедура выполняется для каждого веса от нейрона скрытого слоя к нейрону в выходном слое.

Рассмотрим один нейрон в скрытом слое, предшествующем выходному слою. При проходе вперед этот нейрон передает свой выходной сигнал нейронам в выходном слое через соединяющие их веса.

Во время обучения эти веса функционируют в обратном порядке, пропуская величину δ от выходного слоя назад к скрытому слою. Каждый из этих весов умножается на величину δ нейрона, к которому он присоединен в выходном слое. Величина δ, необходимая для нейрона скрытого слоя, получается суммированием всех таких произведений и умножением на производную сжимающей функции:

$$\delta_{p,j} = OUT_{p,j}\left(1 - OUT_{p,j}\right)\sum_{q}\delta_{q,k}w_{pq,k}. \tag{14}$$

Когда значение δ получено, веса между входным слоем и скрытым слоем j могут быть скорректированы с помощью формул (13) и (14), в которых индексы необходимо модифицировать в соответствии со слоем. То есть процесс обучения представляет собой вычисление δ для каждого нейрона в данном слое и коррекцию всех весов данного слоя.

Для ускорения обучения сети алгоритм обучения модифицируют путем введения нейронного смещения и стабилизирующего множителя, условно называемого "импульсом".

Введение нейронного смещения позволяет сдвигать начало отсчета передаточной функции и по сути является процедурой аналогичной подстройке порога персептронного нейрона. Смещение вводится посредством добавления к каждому слою нейронов дополнительного нейрона, на который подается сигнал, равный +1, а не

выходу нейрона предыдущего слоя. В процессе обучения вес данного нейрона корректируется также как и остальные веса нейронов.

Введение импульса позволяет ускорить обучение сети при использовании алгоритма обратного распространения. Этот метод заключается в добавлении к коррекции веса члена, пропорционального величине предыдущего изменения веса. Как только происходит коррекция, она запоминается и служит для модификации всех последующих коррекций. Уравнение коррекции принимает следующий вид:

$$w_{pq,k}^{(n+1)} = \alpha \quad \Delta w_{pq,k}^{n} + \eta \quad \left(\delta_{q,k} OUT_{pj}\right). \tag{15}$$

Затем вычисляется изменение веса:

$$w_{pq,k}^{n+1} = w_{pq,k}^{n} + \Delta w_{pq,k}^{(n+1)}, \tag{16}$$

где α - коэффициент импульса, обычно ≈ 0.9 [5].

Описанный метод обучения МСП применяется в данной работе.

2.3.3. Применение МСП

В работе МСП применялся для построения модели "чёрного ящика". Пусть есть некоторая система (черный ящик), для которой известен набор входов и соответствующий им набор выходов. В этом случае возможна постановка двух задач.

1. Прямая задача. ИНС применялась для моделирования работы системы. Т.е. на входы МСП подавался сигнал с входа системы, а с выхода снимался сигнал, соответствующий выходному сигналу системы. Польза такого применения состоит в том, что мы получаем оператор, описывающий преобразование сигнала в системе. Этот оператор представлен в виде разложения в нейросетевом базисе.
2. Обратная задача. Здесь по известным выходам системы сеть восстанавливала её входы. Таким образом, МСП был применён для коррекции сигнала ЛФД.

2.4. Сети на основе радиально-базисных функций

2.4.1. Радиальные функции

Радиальные функции – это специальный класс функций, основной характеристикой которых является то, что их отклик монотонно изменяется в зависимости от расстояния до центральной точки. Сеть с радиальными базисными функциями считается нелинейной, если базисная функция может двигаться вдоль осей или изменять размер, или если сеть содержит более одного скрытого слоя. В класс радиальных функций входят функции Гаусса (17а) и Коши (17б), мультиквадратичная (17в) и обратная мультиквадратичная (17г) функции:

$$F(\xi) = \exp\left(-\frac{\xi^2}{2\sigma^2}\right), \qquad (17\ а)$$

$$F(\xi) = \frac{1}{\sigma^2 + \xi^2}, \qquad (17\ б)$$

$$F(\xi) = \sqrt{\xi^2 + \sigma^2}, \qquad (17\ в)$$

$$F(\xi) = \frac{1}{\sqrt{\xi^2 + \sigma^2}}. \qquad (17\ г)$$

Параметр σ определяет радиус влияния каждой базисной функции, то есть, как быстро базисная функция стремиться к нулю с удалением от центра.

2.4.2. Архитектура РБФ-сети

Представленная на рисунке 3 РБФ-сеть состоит из двух слоев: скрытый слой радиальных базисных нейронов и выходной слой с классической передаточной функцией и классическими нейронами. [4, 26] Когда мы подаем входной вектор на такую сеть, каждый нейрон в радиальном базисном слое выдает на выходе величину соответствия входного вектора и вектора центров нейрона, которая близка к нулю в случае их различия, и тогда не оказывает практически никакого влияния на выходной слой. Наиболее оптимальной для точной аппроксимации оказывается такая структура сети, где число нейронов скрытого слоя равно числу входных данных. В таком случае центры нейронов скрытого слоя практически совпадают со значениями входных векторов. Данная структура РБФ-сети является стандартной, то есть сети с более сложной структурой не рассматриваются.

Архитектура РБФ-сети

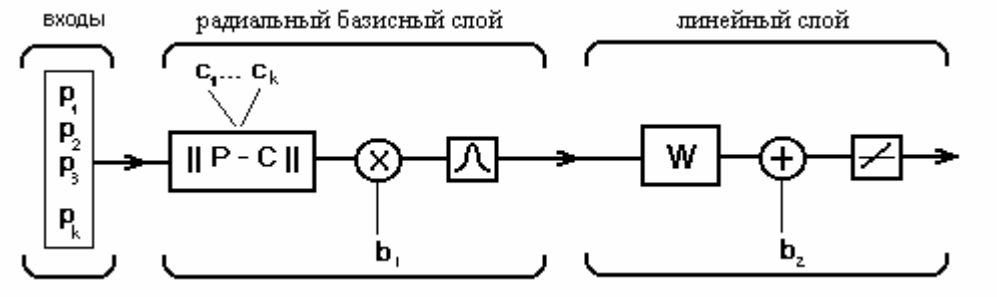


Рис. 3

Нейрон РБФ-слоя

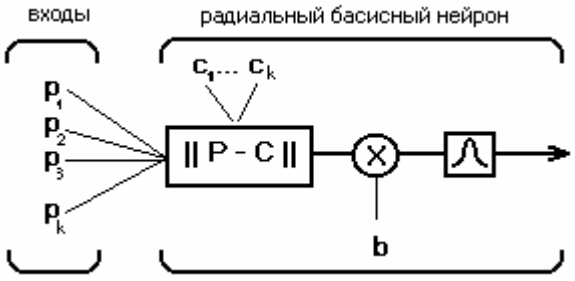


Рис.4

2.4.3. Модель нейрона РБФ-сети

На рисунке 4. приведена структура нейрона, расположенного в РБФ-слое сети. Подающиеся на вход сети данные из внешней среды, поступают к нейрону вместе с вектором, определяющим центр базисной функции. От найденного модуля разности $\|P-C\|$, умноженного на смещение b, вычисляется передаточная радиальная базисная функция, сигнал на выходе которой и является выходом нейрона [24, 25].

С уменьшением расстояния между входным вектором и вектором, определяющим центр базисной функции, выход нейрона увеличивается. Таким образом, радиальный базисный нейрон работает как детектор, то есть на его выходе будет максимальное значение, когда вход P эквивалентен вектору C. Смещение b компенсирует различие в средних значениях от входа и выхода нейрона.

2.4.4. Обучение РБФ-сети

Ключевой особенностью РБФ-сети является различие функций первого и второго слоя сети. Весовые коэффициенты первого слоя (т.е. параметры базисных функций) определяются в результате применения методов обучения без учителя. Это приводит к двухэтапному процессу обучения. На первом этапе определяется смещение и положение центров базисных функций. На втором этапе определяются весовые коэффициенты следующего слоя с помощью стандартных методов с учителем. За счет применения методики обучения без учителя на первом этапе значительно ускоряется процесс обучения сети по сравнению с сетями классической структуры [24,26].

Формальное доказательство того, что линейная суперпозиция радиальных функций пригодна для универсальной аппроксимации, получено для сетей с гауссовской функцией, в которой ширина функций является настраиваемым параметром [25]. В той же работе было показано, что многослойные персептроны не обладают таким свойством. Однако полученные результаты не давали процедур построения сетей. В дальнейшем было доказано, что РБФ-сети обладают свойством наилучшей аппроксимации (то есть в наборе аппроксимирующих функций существует функция, которая обеспечивает минимум ошибки аппроксимации для любой аппроксимируемой функции).

Одной из основных особенностей РБФ-сетей является возможность их аналитического описания. Далее в работе предлагается метод аналитического

расчета параметров нейронной сети, позволяющий существенно сократить временные затраты на обучение сети при сохранении требуемой точности.

Далее приведены сравнительные результаты аппроксимации выходных сигналов кремниевых генераторных диодов нейронными сетями с различными радиальными базисными функциями. На рисунках сплошной линией обозначен входной набор, пунктирной– аппроксимация.

На рисунке 5 приведена аппроксимация гауссовой функцией с шириной 0.01. Процедура потребовала 16 итераций.

На рисунке 6 представлена аппроксимация функцией Коши с шириной, равной 0.04. Процедура оптимизации потребовала 16 итераций. При ширине функции, равной 0.01 сети потребовалось на оптимизацию 20 итераций.

На рисунке 7 приведена аппроксимация мультиквадратичной функцией с шириной, равной 0.02 (*а*) и обратной мультиквадратичной функцией (*б*). Процедура оптимизации потребовала 10 и 14 итераций соответственно.

Результаты исследований показывают, что применение различных базисных функций требует тщательного подбора ширины данной функции. Правильно подобранная ширина позволяет несколько уменьшить число итераций, необходимых для оптимизации сети; в любом случае точность аппроксимации практически не меняется.

Проведенные исследования показали применимость использования основных радиальных базисных функций в качестве базиса при аппроксимации выходных сигналов кремниевых генераторных диодов нейронными сетями.

Аппроксимация гауссовой функцией

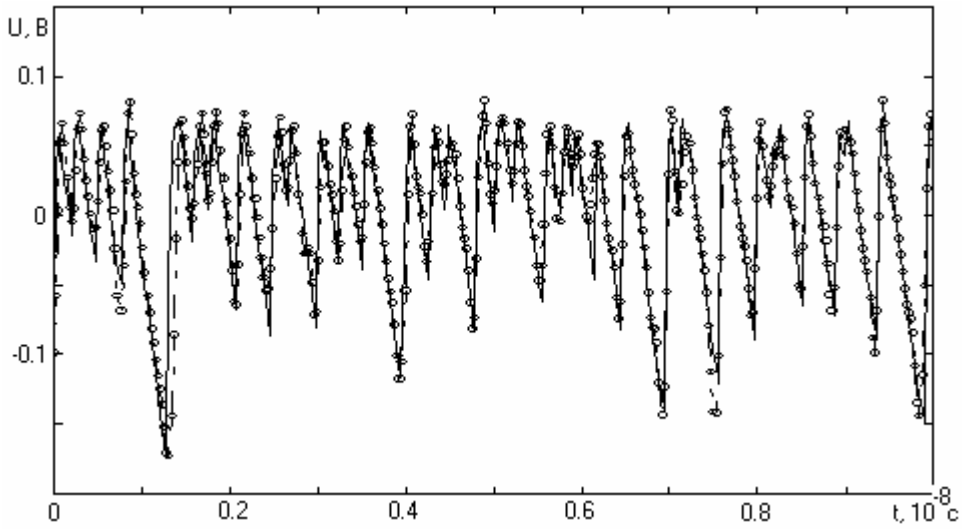


Рис. 5

Аппроксимация функцией Коши

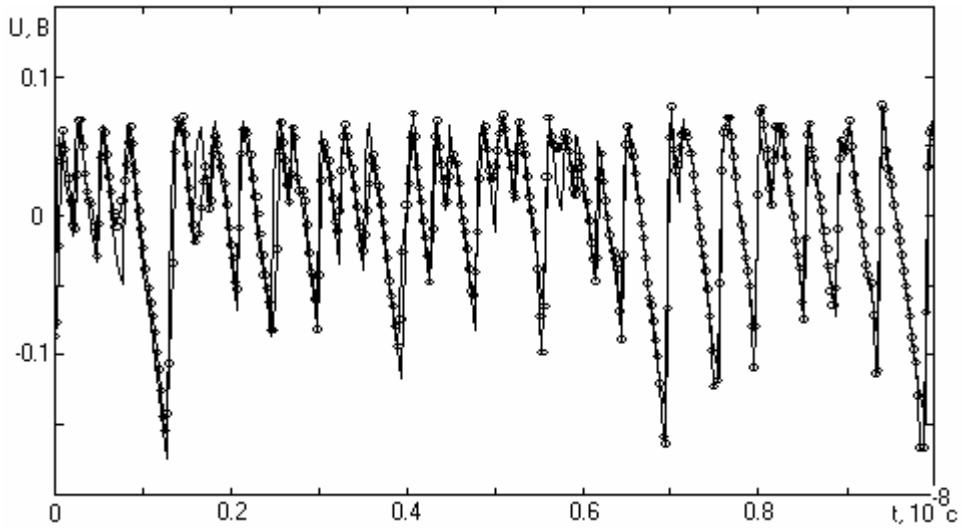
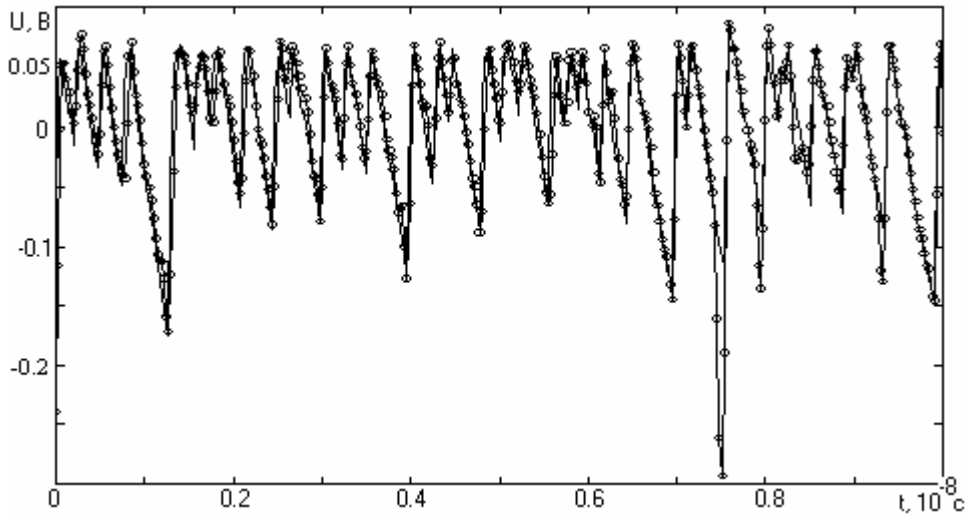
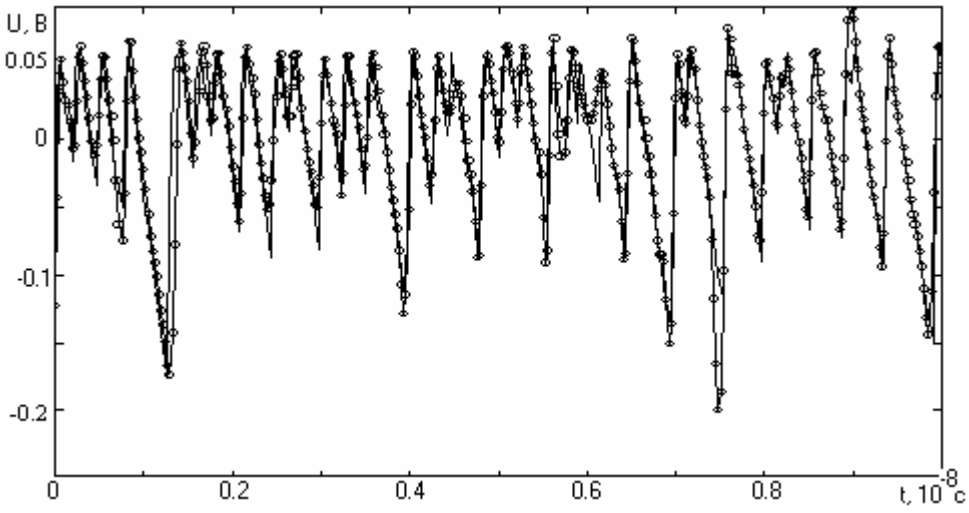


Рис. 6

Аппроксимация мультиквадратичной (а) и
обратной мультиквадратичной (б) функциями



а)



б)

Рис. 7

2.4.5. Расчет и оптимизация РБФ-сети

2.4.5.1. Постановка задачи и расчет весов, минимизирующих целевую функцию

В общем случае задача процесса обучения заключается в минимизации среднеквадратичной ошибки:

$$S = \sum_{i=1}^{p} (y_i - f(x_i))^2 , \tag{18}$$

где (x_i,y_i) – набор обучающих пар, а $f(x_i) = \sum_{j=1}^{m} w_j * h_j(x)$, m-число весов, p-число входов сети [25].

Если в результате решения задачи необходимо получить гладкую аппроксимирующую функцию, тогда к среднеквадратичной ошибке добавляется член, регулирующий гладкость функции:

$$\tilde{E} = E + \lambda * \Omega , \tag{19}$$

где E – среднеквадратичная ошибка, λ - параметр регуляризации, а Ω - и есть дополнительное слагаемое. В зависимости от вида Ω существуют различные методы. В частности, $\Omega = \frac{1}{2}\sum_j w_j^2$ используется в методе гребенчатой регрессии (ridge regression), который и будет рассматриваться в данной работе. В таком случае целью задачи является минимизация целевой функции вида:

$$C = \sum_{i=1}^{p} (y_i - f(x_i))^2 - \sum_{j=1}^{m} \lambda_j w_j^2 . \tag{20}$$

Для минимизации данной функции необходимо приравнять к нулю ее первую производную, что приводит к системе m уравнений вида:

$$\sum_{i=1}^{p} f(x_i) * h_j(x_i) + \lambda_j * w_j = \sum_{i=1}^{p} y_i * h_j(x_i), \; j = \overline{1,m} , \tag{21}$$

или с использованием матричных обозначений:

$$h_j^T * f + \lambda_j * w_j = h_j^T * y,\ j = \overline{1,m} \tag{22}$$

$$H^T * f + \Lambda * w = H^T * y, \tag{23}$$

где Λ - диагональная матрица с диагональными элементами λ_j, H – матрица плана (design matrix). Учитывая, что $f_i = \overline{h}_i^T w$, где $\overline{h}_i = \begin{bmatrix} h_1(x_i) \\ ..... \\ h_m(x_i) \end{bmatrix}$, получаем:

$$w = (H^T * H + \Lambda)^{-1} * H^T * y\ . \tag{24}$$

Выражение (24) является формулой для расчета весовых коэффициентов, при которых целевая функция будет иметь минимум. На (24) ссылаются как на систему нормальных уравнений.

Матрица $A^{-1} = (H^T * H + \Lambda)^{-1}$ получила название матрицы отклонений. Подставив ее в систему (24) получим:

$$w = A^{-1} * H^T * y\,. \tag{25}$$

Следует учитывать, что на данном этапе нам неизвестно оптимальное значение параметра регуляризации. Его значение выбирают в зависимости от исходных данных, на основании которых решается задача аппроксимации.

2.4.5.2. Оптимальное значение целевой функции

При анализе нейронных сетей с использованием методов линейной алгебры часто оказывается выгодным использовать матрицу проекций (упрощается вид аналитических выражений и облегчается анализ ошибки):

$$P = I_p - H * A^{-1} * H^T\,. \tag{26}$$

Эта матрица проектирует вектора из *p*-размерного пространства входов в *m*-размерное пространство весов и позволяет легко оценить полученную ошибку $P * y$.

Так как, по определению $f = H * w = H * A^{-1} * H^T * y$, то:

$$y - f = P * y \ . \tag{27}$$

Отсюда следует, что оптимальная среднеквадратичная ошибка будет равна:

$$S = (y - f)^T * (y - f) = y^T * P^T * P * y = y^T * P^2 * y \, . \tag{28}$$

Аналогично, оптимальная целевая функция:

$$\begin{aligned} C &= (H * w - y)^T * (H * w - y) + w^T * \Lambda * w = \\ &= y^T * P^2 * y + y^T * (P - P^2) * y = y^T * P * y \end{aligned} \, . \tag{29}$$

2.4.5.3. Оптимальное число весовых коэффициентов

По аналогии с формулой для оценки ошибки $\sigma^2 = \frac{1}{p-1} * \sum_{i=1}^{p} (x_i - \bar{x})^2$ в теории нейронных сетей вводится: $\sigma^2 = \frac{S}{p-m}$. Однако при использовании гребенчатой регрессии вводится понятие эффективного числа параметров γ такого, что:

$$\gamma = p - trace(P) \, , \tag{30}$$

$$\sigma^2 = \frac{S}{p-\gamma} \, . \tag{31}$$

Если не использовать методы регуляризации, то $\gamma=m$.

2.4.5.4. Критерии выбора модели системы

предназначены для получения оценок ошибки предсказания, то есть оценки, насколько хорошо обученная сеть будет работать с новыми данными.

Одним из наиболее широко используемых методов получения оценки является метод перекрестной проверки по сокращенным выборкам (leave-one-out cross-validation), используемый, когда данных недостаточно и для настройки и для тестирования сети. Тогда набор обучающих пар можно представить в виде 2 частей: одна часть для обучения и одна часть для тестирования. Ошибка считается по всем возможным вариантам разбиения по части обучающих пар для тестирования. В предельном варианте из набора в p пар только 1 пара выделяется для тестирования, и среднеквадратичная ошибка считается по p возможным комбинациям. Преимуществом данного метода является то, что все данные могут быть использованы для обучения.

Достоинством РБФ-сетей является то, что оценка может быть подсчитана аналитически:

$$\sigma^2_{LOO} = \frac{y^T * P * (diag(P))^{-2} * P * y}{p}. \quad (32)$$

Другие методы получения оценки:

1. Метод обобщенной перекрестной проверки GCV (generalised cross-validation):

$$\sigma^2_{GCV} = \frac{p * y^T * P^2 * y}{(p-\gamma)^2}. \quad (33)$$

2. Метод несмещенной оценки вариации UEV (unbiased estimate of variance):

$$\sigma^2_{UEV} = \frac{y^T * P^2 * y}{p-\gamma}. \quad (34)$$

3. Метод конечной ошибки предсказания FPE (final prediction error):

$$\sigma^2_{FPE} = \frac{(p+\gamma) * y^T * P^2 * y}{(p-\gamma) * p}. \quad (35)$$

4. Метод байессовского информационного критерия BIC (Bayesian information criterion):

$$\sigma_{BIC}^2 = \frac{p + (\ln p - 1) * \gamma}{p - \gamma} * \frac{y^T * P^2 * y}{p}. \quad (36)$$

Методы UEV, FPE, GCV и BIC могут быть представлены в общем виде:

$$\sigma_{xyz}^2 = \frac{\Gamma_{xyz} * y^T * P^2 * y}{p}. \quad (37)$$

где $\Gamma_{UEV} \le \Gamma_{FPE} \le \Gamma_{GCV} \le \Gamma_{BIC}$ и

$$\begin{aligned} \Gamma_{UEV} &= \frac{p}{p - \gamma} \quad \Gamma_{FPE} = \frac{p + \gamma}{p - \gamma} \\ \Gamma_{GCV} &= \frac{p^2}{(p - \gamma)^2} \quad \Gamma_{BIC} = \frac{p + (\ln p - 1) * \gamma}{p - \gamma} \end{aligned}. \quad (38)$$

2.4.5.5. Метод гребенчатой регрессии

Если на вход сети подается вектор *X*, на выходе получаем *f(x)*, а значения требуемых выходов *y(x)*, то среднеквадратичная ошибка по всем элементам (*MSE*): $MSE=<(y(x) - f(x))^2>$. Она представляется в виде 2 частей:

$$MSE = (y(x) - < f(x) >)^2 + < (f(x) - < f(x) >)^2 > . \quad (39)$$

Первое слагаемое получило название *смещение*, второе – *отклонение (variance)*. При таком подходе более простой структуре сети соответствуют меньшие смещения (различие средних значений), а отклонение отвечает за гибкость системы (чувствительность сети к отдельным наборам данных) [26].

Метод гребенчатой регрессии позволяет за счет явного введения смещений уменьшить эффективное число параметров, однако уменьшается диапазон значений функции и, следовательно, гибкость сети.

Наиболее удобный метод контроля соотношения гибкости сети и малого числа параметров заключается в добавлении к целевой функции дополнительного члена – (19) - (20). Значения весовых коэффициентов рассчитываются по (25).

2.4.5.6. Выбор параметра регуляризации

Методы выбора модели системы позволяют также оценить значение оптимального параметра регуляризации. Например, используя метод GCV, получим:

$$\sigma_{GCV}^2 = \frac{p * y^T * P^2 * y}{(trace(P))^2}. \tag{40}$$

Для того, чтобы найти минимум этой ошибки как функцию параметра регуляризации, необходимо продифференцировать это уравнение по λ и приравнять производную к 0. Можно показать, что:

$$\frac{\partial P}{\partial \lambda} = -H * \frac{\partial A^{-1}}{\partial \lambda} * H^T = H * A^{-2} * H^T, \tag{41}$$

$$\frac{\partial}{\partial \lambda} trace(P) = -trace(H * \frac{\partial A^{-1}}{\partial \lambda} * H^T) = trace(A^{-1} - \lambda * A^{-2}), \tag{42}$$

$$\frac{\partial}{\partial \lambda} y^T * P^2 * y = 2 * \lambda * w^T * A^{-1} * w. \tag{43}$$

Подставив выражения (41) - (43) в (40), получим:

$$\lambda * w^T * A^{-1} * w * trace(P) = y^T * P^2 * y * trace(A^{-1} - \lambda * A^{-2}). \tag{44}$$

После упрощения данного выражения получаем рекуррентную формулу:

$$\lambda = \frac{y^T * P^2 * y * trace(A^{-1} - \lambda A^{-2})}{w^T * A^{-1} * w * (p - \gamma)}. \tag{45}$$

Другие методы выбора модели системы дают следующие выражения:

$$\text{(UEV):}\ \lambda = \frac{y^T * P^2 * y * trace(A^{-1} - \lambda * A^{-2})}{2 * w^T * A^{-1} * w * (p - \gamma)}, \tag{46}$$

$$\text{(FPE):}\quad \lambda = \frac{y^T * P^2 * y * trace(A^{-1} - \lambda * A^{-2}) * p}{w^T * A^{-1} * w * (p-\gamma) * (p+\gamma)}, \tag{47}$$

$$\text{(BIC):}\quad \lambda = \frac{y^T * P^2 * y * trace(A^{-1} - \lambda * A^{-2}) p * \log p}{2 * w^T * A^{-1} * w * (p-\gamma) * (p+\gamma * (\log p - 1))}. \tag{48}$$

2.4.5.7. Изменение числа базисных функций и обучающих пар

Одна из проблем при подборе оптимального числа параметров заключается в нахождении эффекта от добавления или исключения базисной функции, добавления или исключения обучающей пары. Конечно, такой эффект можно просчитать простым переобучением сети, однако временные затраты в таком случае будут достаточно большими. В случае РБФ-сети оказывается возможным с помощью аналитических формул определить эффект от работы сети в таких ситуациях.

При расчете вышеперечисленных эффектов основную сложность представляет пересчет матриц A^{-1} и P. При пересчете обратной матрицы используется формула:

Если $A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$, то

$$A^{-1} = \begin{bmatrix} (A_1 - A_2 * A_4^{-1} A_3)^{-1} & A_1^{-1} * A_2 * (A_3 * A_1^{-1} A_2 - A_4)^{-1} \\ (A_3 * A_1^{-1} A_2 - A_4)^{-1} * A_3 * A_1^{-1} & (A_4 - A_3 * A_1^{-1} A_2)^{-1} \end{bmatrix}, \tag{49}$$

или в более простой форме:

$$A^{-1} = \begin{bmatrix} A_1^{-1} + A_1^{-1} * A_2 * \Delta^{-1} * A_3 * A_1^{-1} & -A_1^{-1} * A_2 * \Delta^{-1} \\ -\Delta^{-1} * A_3 * A_1^{-1} & \Delta^{-1} \end{bmatrix}, \tag{50}$$

где $\Delta = A_4 - A_3 * A_1^{-1} * A_2$.

Число операций, необходимых для пересчета матрицы P, приведено в таблице 1.

Таблица 1.

Операция	Полный пересчет	Пересчет по формулам
Добавление функции	$m^3+pm^2+p^2m$	p^2
Удаление функции	$m^3+pm^2+p^2m$	p^2
Добавление обуч. пары	$m^3+pm^2+p^2m$	$2m^2+pm+p^2$
Удаление обуч. Пары	$m^3+pm^2+p^2m$	$2m^2+pm+p^2$

Добавление новой базисной функции равносильно добавлению нового столбца в матрицу H: $H_{m+1} = \begin{bmatrix} H_m & h_{m+1} \end{bmatrix}$; новая матрица отклонений:

$$A_{m+1} = H_{m+1}^T * H_{m+1} + \Lambda_{m+1} = \begin{bmatrix} A_m & H_m^T * h_{m+1} \\ h_{m+1}^T * H_m & \lambda_{m+1} + h_{m+1}^T * h_{m+1} \end{bmatrix},$$

используя формулу (42), получим новую обратную матрицу и матрицу проекций:

$$A_{m+1}^{-1} = \begin{bmatrix} A_m^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\lambda_{m+1} + h_{m+1}^T * P_m * h_{m+1}} \begin{bmatrix} A_m^{-1} * H^T * h_{m-1} \\ -1 \end{bmatrix} \begin{bmatrix} A_m^{-1} * H^T * h_{m-1} \\ -1 \end{bmatrix}^T, \quad (51)$$

$$P_{m+1} = P_m - \frac{P_m * h_{m+1} * h_{m+1}^T * P_m}{\lambda_{m+1} + h_{m+1}^T * P_m * h_{m+1}}, \quad (52)$$

Аналогично рассуждениям, проведенным в предыдущем пункте, получим матрицу проекций с исключенным *j* столбцом:

$$P_{m-1} = P_m + \frac{P_m * h_j * h_j^T * P_m}{\lambda_j + h_j^T * P_m * h_j}. \quad (53)$$

В случае изменения числа базисных функций матрица проекций сохраняет свою размерность ($p*p$), изменяется размерность матрицы отклонений.

В случае изменения числа обучающих пар меняется размерность матрицы проекций. Добавление обучающей пары равносильно добавлению строки в матрицу плана:

$$H_{p+1} = \begin{bmatrix} H_p \\ h_{p+1}^T \end{bmatrix}, \quad h_{p+1}^T = \begin{bmatrix} h_1(x_{p+1}) & ..... & h_m(x_{p+1}) \end{bmatrix}. \quad (54)$$

В этом случае новая матрица отклонений и матрица проекций:

$$A_{p+1} = H_{p+1}^T * H_{p+1} = A_p + h_{p+1} * h_{p+1}^T, \quad (55)$$

$$A_{p+1}^{-1} = A_p^{-1} - \frac{A_p^{-1} * h_{p+1} * h_{p+1}^T * A_p^{-1}}{1 + h_{p+1}^{-1} * A_p^{-1} * h_{p+1}}, \quad (56)$$

$$P_{p+1} = \begin{bmatrix} P_p^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{1 + h_{p+1}^T * A_p^{-1} * h_{p+1}} * \begin{bmatrix} H_p * A_p^{-1} * h_{p+1} \\ -1 \end{bmatrix} \begin{bmatrix} H_p * A_p^{-1} * h_{p+1} \\ -1 \end{bmatrix}^T. \quad (57)$$

Аналогично рассуждениям предыдущего пункта можно получить:

$$A_{p-1}^{-1} = A_p^{-1} + \frac{A_p^{-1} * h_i * h_i^T * A_p^{-1}}{1 - h_i^{-1} * A_p^{-1} * h_i}. \quad (58)$$

2.5. Выводы

Таким образом, в этом разделе были рассмотрены основы построения, функционировании и применения ИНС для моделирования систем и обработки сигналов. Был предложен метод аналитического расчета параметров РБФ-сети, позволяющий избежать излишних временных затрат на обучение сети, которые становятся довольно значительными при большом объеме входных данных даже в случае РБФ-сети; на основе данного метода можно также просчитать эффект от изменения числа базисных функций и обучающих пар. Описана методика оптимизации сети, позволяющая рассчитать оптимальные параметры РБФ-сети, при которых точность аппроксимации сравнима с точностью точной аппроксимации.

3. МОДЕЛИРОВАНИЕ И ВОССТАНОВЛЕНИЕ ХАРАКТЕРИСТИК КРЕМНИЕВЫХ ДИОДОВ

3.1. Введение

В данном разделе в качестве исходных данных для настройки нейросетевых моделей используются шумовые характеристики кремниевых генераторных диодов [27], работающих в режиме микроплазменного пробоя *p*-*n*-перехода [12]. В первом пункте приводятся описание физического эксперимента и полученные результаты измерений. Во втором подразделе производится моделирование характеристик диодов, разрабатывается алгоритм для восстановления параметров работы приборов по выходным спектральным характеристикам. В третьем пункте с помощью нейросетевых моделей проводится анализ возможности прогнозирования и аппроксимации выходных шумовых характеристик диодов путем построения их импульсных характеристик.

3.2. Шумовые характеристики кремниевых диодов

Диоды кремниевые генераторные применяются в качестве источника шумовых сигналов в калибровочных системах, а также в системах кодирования и защиты информации. Такие полупроводниковые приборы имеют равномерную спектральную плотность шума в заданном интервале частот. Однако воздействие внешних факторов – температуры, электромагнитных полей или флуктуаций режимов работы приборов может негативно сказываться на их выходных характеристиках, приводя к ошибкам и снижению надежности аппаратуры. Поэтому был проведен физический эксперимент по исследованию влияния режимов работы кремниевых диодов на их выходные характеристики.

Целью исследований являлась оценка влияния сопротивления нагрузки и величины приложенного напряжения на форму выходных импульсов и спектральную плотность шумов.

На рисунке 8 представлены примеры шумовых импульсов для двух значений напряжения питания и одинаковом сопротивлении нагрузки. Как видно из рисунке 8, при повышении напряжения наблюдается изменение формы импульсов. При этом увеличивается частота следования импульсов и их амплитуда. Эксперимент проводился для четырех значений сопротивлений нагрузки: 2.8 кОм, 5.6 кОм, 20 кОм и 31.2 кОм. Микроплазменный пробой или нестабильность тока через диод наблюдались в

некотором интервале напряжений питания, зависящим от сопротивления нагрузки: от 7.7 В до 8.20 В (2.8 кОм), 8.50 В (5.6 кОм), 10.50 В (20 кОм), 11.80 В (31.2 кОм).

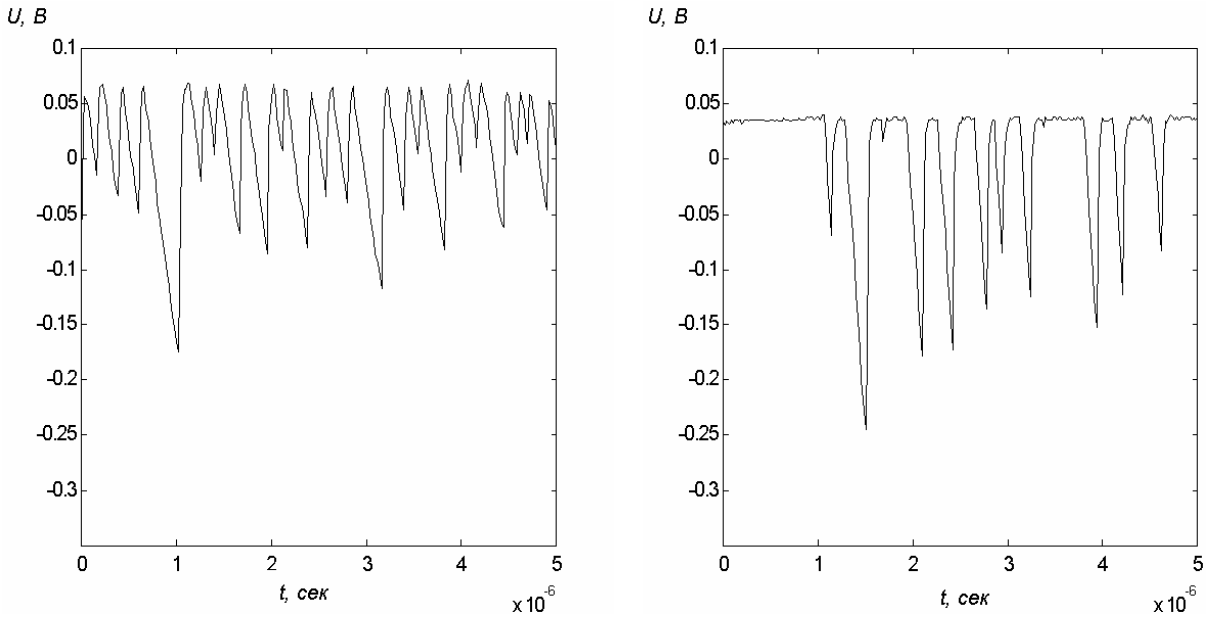
Измеренные шумовые импульсы диодов подвергались математической обработке с использованием преобразования Фурье, после чего рассчитывалась спектральная плотность мощности. Погрешность расчета спектральной плотности мощности не превышала 7 %.

На рисунке 9 представлены изменение спектральной плотности мощности выходных шумовых импульсов в зависимости от напряжения питания для различных сопротивлений нагрузки. На графиках хорошо просматриваются эти области нестабильности – кривые, имеющие максимум. При этом положение максимума менялось с изменением величины напряжения питания и сопротивления нагрузки. Так, до некоторого напряжения $U_{кр}$ положение максимума смещалось в область более высоких частот (увеличение частоты шумовых импульсов), при напряжениях, больших $U_{кр}$, максимум смещается обратно в область более низких частот (увеличивается длительность импульсов).

Полученные временные и спектральные характеристики послужили экспериментальным материалом для построения моделей и методов анализа изучаемых генераторных диодов.

Шумовые импульсы генераторного диода:

а) напряжение питания 8.80 В; б) напряжение питания 9.90 В.



а) б)

Рис. 8

Изменение спектральной плотности мощности генераторного диода для различных сопротивлений нагрузки в зависимости от напряжения питания:
а) 2.8 кОм; б) 5.6 кОм; в) 20 кОм; г) 31.2 кОм.

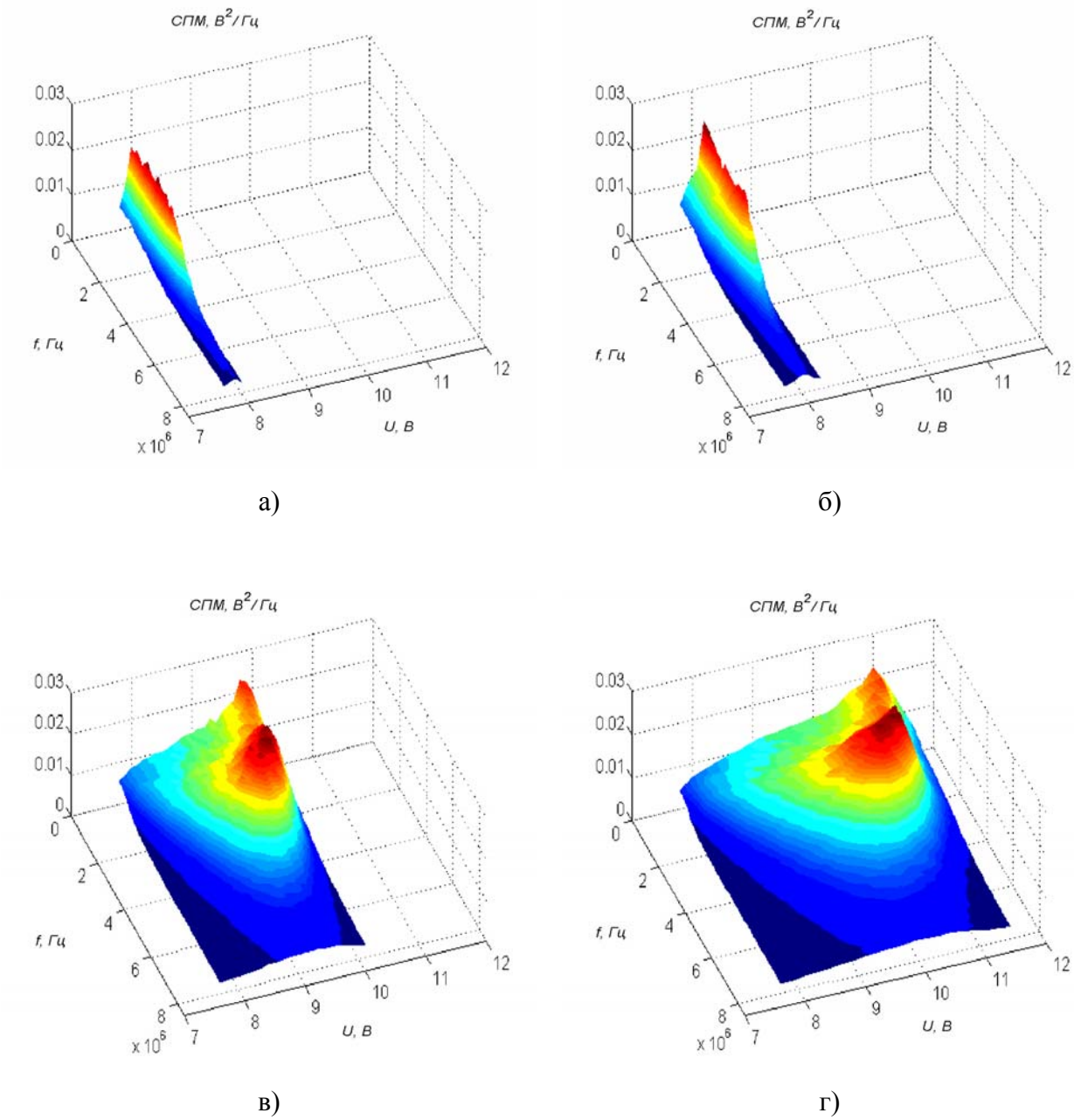


Рис. 9

3.3. Моделирование характеристик шумовых диодов

Исследуем возможность использования ИНС для восстановления параметров режима работы диодов по выходным данным.

3.3.1. Постановка задачи

Даны шумовые спектры, снятые с диода при наборе значений входного сопротивления (R) и напряжений питания (V). Эти спектры представлены на рисунке 9. Надо по шумовым спектрам восстановить параметр эксперимента R.

Важным этапом является предварительная обработка шумовых спектров. Поскольку различие между первой гармоникой спектра и остальными гармониками может составлять несколько порядков, необходимо масштабировать и спектр и его первую гармонику таким образом, чтобы они находились в интервале 10^{-1}–10^{2}. Кроме того, нет смысла работать с высокочастотной областью спектра, так как она малоинформативна. Поэтому при работе со спектрами ограничивались первыми 32 отсчётами. Таким образом, на вход ИНС подавалось 32 спектральных отсчёта.

3.3.2. Выбор структуры сети

В качестве базовой структуры сети был выбран МСП с 32 входами и 1 выходом. Для выявления оптимальных размеров скрытых слоёв проводился вычислительный эксперимент, результатом которого явилась зависимость ошибки после обучения от количества нейронов в двух скрытых слоях см. рисунок 10. По вертикали отложено количество нейронов в первом скрытом слое (от 1 до 16), по горизонтали – во втором (от 1 до 8).

Исходя из результатов эксперимента, в качестве оптимальной структуры был выбран МСП с размерами 32x8x4x1, представленный на рисунке 11.

Зависимость ошибки сети после 2000 итераций обучения от размеров скрытых слоёв.

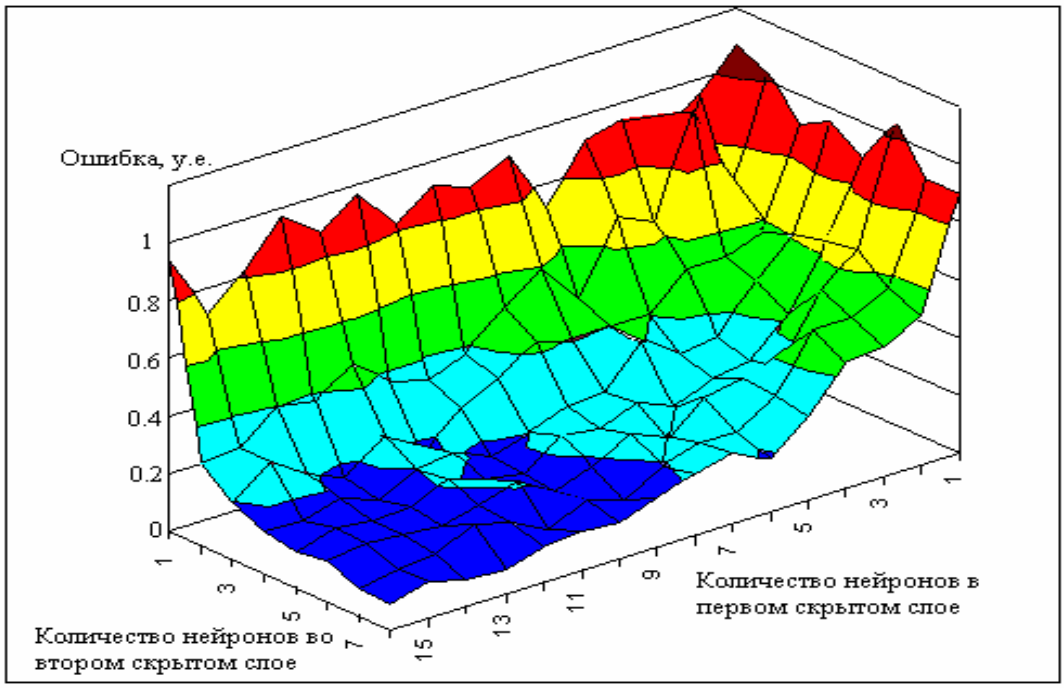


Рис.10

Трёхслойный персептрон для восстановления параметров эксперимента.

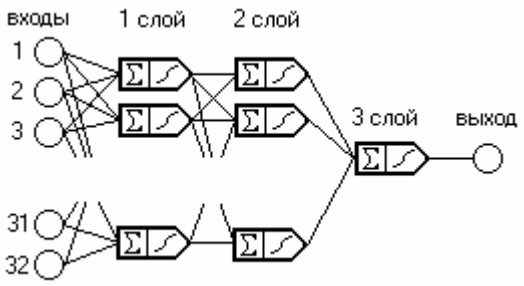


Рис.11

3.3.3. Обучение по всей совокупности

Чтобы убедиться в возможности использования ИНС для задач такого рода, было проведено обучение сети на всей совокупности экспериментальных данных (56 спектров). Результат обучения в течение 10000 итераций представлен на рисунке 12.

Из рисунке 12 видно, что восстановленное сопротивление очень близко к экспериментальному, причём разница между ними убывает с увеличением числа итераций. Это позволяет зделать вывод о том, что при достаточно большом числе итераций можно восстановить параметр эксперимента с заданной точностью.

3.3.4. Обучение по неполной совокупности

Для того, чтобы проверить способность сети восстанавливать значения параметра в тех точках, для которых экспериментальных данных нет, проводилось обучение на 37 спектрах, что составляет 2/3, и 28 спектра – 1/2 всей совокупности. Результаты представлены на рисунке 13.

Видно, что в случае (а) количество обучающих пар ещё не достаточно для правильного обучения сети, тогда как в случае (б) сеть даёт удовлетворительный результат.

Таким образом, было показано, что МСП в состоянии решать задачу восстановления экспериментальных параметров по выходным шумовым спектрам.

Результаты восстановления сопротивления нагрузки (R) нейронной сетью: линией показано реальное R, кружками – восстановленное

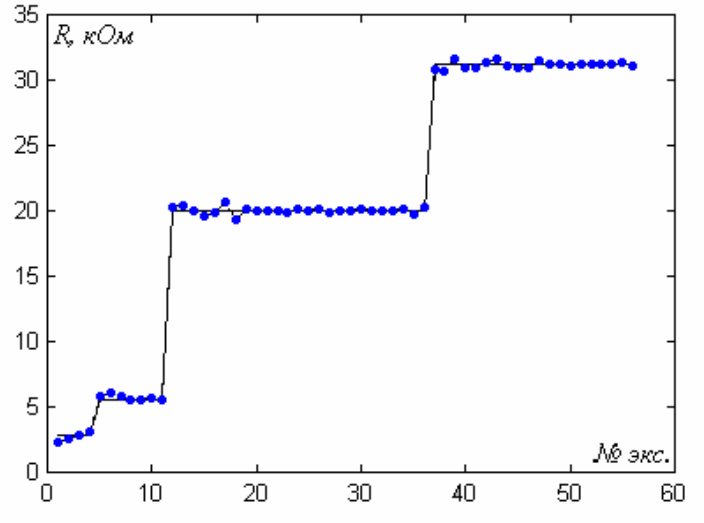
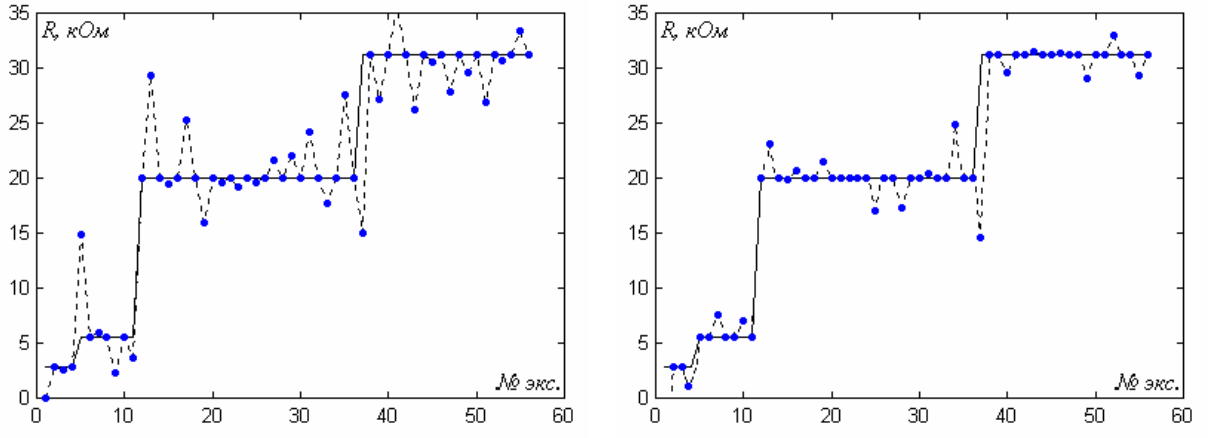


Рис.12

Восстановление параметра в при обучении на неполной совокупности. Обучающими являются: а) 1/2 всех спектров, б) 2/3 всех спектров



а) б)

Рис.13

3.4. Аппроксимация и прогнозирование выходных данных

В данном пункте приведены результаты аппроксимации и прогнозирования шумовых выходных сигналов кремниевых диодов РБФ-сетями различной длины. Использовался прямой расчет весовых коэффициентов и оптимизация параметров сети на основе метода, описанного в пункте 2.3. Приведены графики зависимости ошибки аппроксимации от параметра регуляризации.

3.4.1. Аппроксимация шумовых выходных сигналов кремниевых диодов

Результат аппроксимации с рассчитанной матрицей плана и рассчитанными весовыми коэффициентами по формуле (25) на основе 200 точек представлен на рисунке 14.

Получение достаточно большой ошибки после расчета или обучения нейронной сети является типичной ситуацией, что доказывает необходимость применения методик оптимизации. Вообще говоря, ошибка аппроксимации может быть несколько уменьшена путем подбора ширины базисной функции, однако на данном этапе невозможно аналитически рассчитать оптимальное значение ширины базисной функции. Поэтому вместо проведения большого числа экспериментов для подбора ширины базисной функции оказывается более выгодным с точки зрения временных затрат использовать оптимизирующие методики (точность аппроксимации во втором случае также оказывается более высокой).

На рисунке 15 представлена аппроксимация после оптимизации параметра регуляризации по рекуррентной формуле критерия BIC с начальным значением, равным 1 (синяя линия – входной набор, зеленая - аппроксимация).

Ниже приведен ход процедуры оптимизации параметра регуляризации:

Таблица 2.

Итерация	λ	BIC	Отн. изменения
0	1.000e+000	1.061e+004	-
9	2.833e+000	2.154e-002	1547
10	2.755e-001	2.153e-002	4007
11	2.684e-001	2.153e-002	5202

Первый столбец содержит номер итерации, второй – значение параметра регуляризации на данной итерации, третий – оценка полученной ошибки по формуле для данного критерия и четвертый столбец – относительные изменения ошибки.

Пример неудачной аппроксимации РБФ-сетью до оптимизации параметра регуляризации (сплошная линия – входной набор, кружки – аппроксимация)

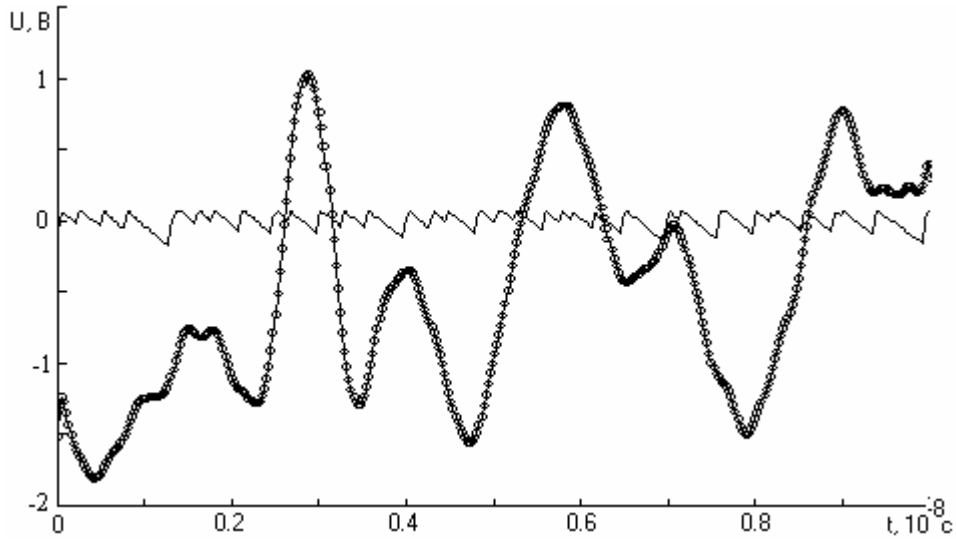


Рис. 14

Аппроксимация после оптимизации параметра регуляризации

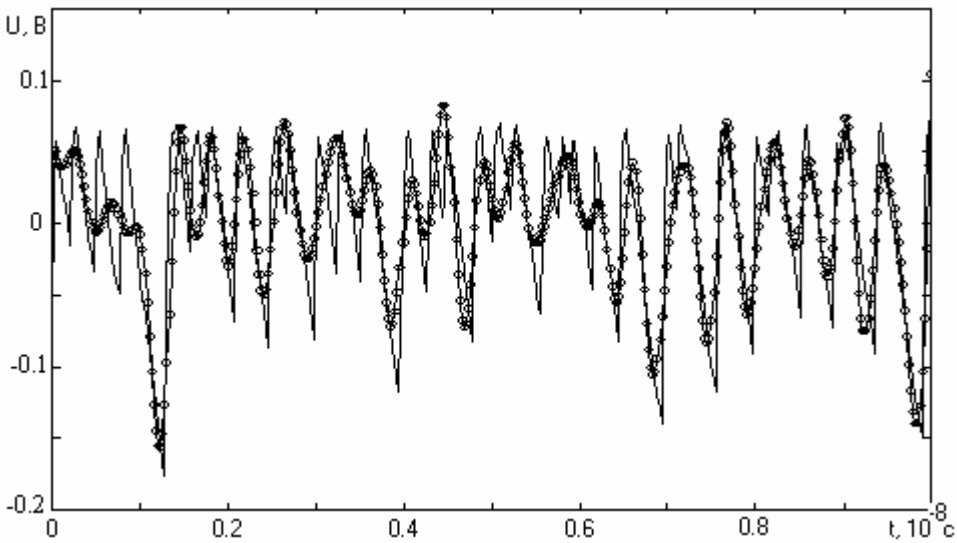


Рис. 15

При аппроксимации РБФ-сетью с использованием других базисных функций число итераций в среднем не меняется (при правильной ширине базисной функции).

Результат аппроксимации с рассчитанной матрицей плана и рассчитанными весовыми коэффициентами на основе 400 точек приведён на рисунке 16-17.

На рисунке 18 представлена зависимость ошибок аппроксимации (e) от параметра регуляризации (λ). Линия 1 соответствует ошибки по критерию BIC, линия 2 – LOO, линия 3 – GCV, линия 5 – UEV, линия 6 – FPE.

Как видно из рисунке 18, в данном примере минимуму среднеквадратичной ошибки соответствуют значения параметра регуляризации, полученные по критериям BIC и LOO, однако результаты исследований показывают, что значения параметра регуляризации, рассчитанные по другим критериям, незначительно отличаются от оптимального. Вообще говоря, для разных задач оптимальное значение параметра регуляризации может рассчитываться по разным критериям, однако погрешность параметра, рассчитанного по другому критерию (и, следовательно, результирующая точность аппроксимации) является незначительной.

Значения оптимального параметра регуляризации, рассчитанные по различным критериям, для примера с 200 точками:

BIC: $\lambda = 2.684e\text{-}001$;

GCV: $\lambda = 2.613e\text{-}001$;

UEV: $\lambda = 2.314e\text{-}001$.

Результат аппроксимации с рассчитанной матрицей плана и рассчитанными весовыми коэффициентами на основе 800 точек показан на рисунке 19 (сплошная линия – входной набор, кружки - аппроксимация).

На рисунке 20 показана зависимость ошибок аппроксимации (e) от параметра регуляризации (λ) (линия 1 соответствует ошибки по критерию BIC, линия 2 – LOO, линия 3 – GCV, линия 4 – FPE, линия 5 – UEV).

Аппроксимирующая функция, полученная на основе
прямого расчета матрицы плана и весовых коэффициентов

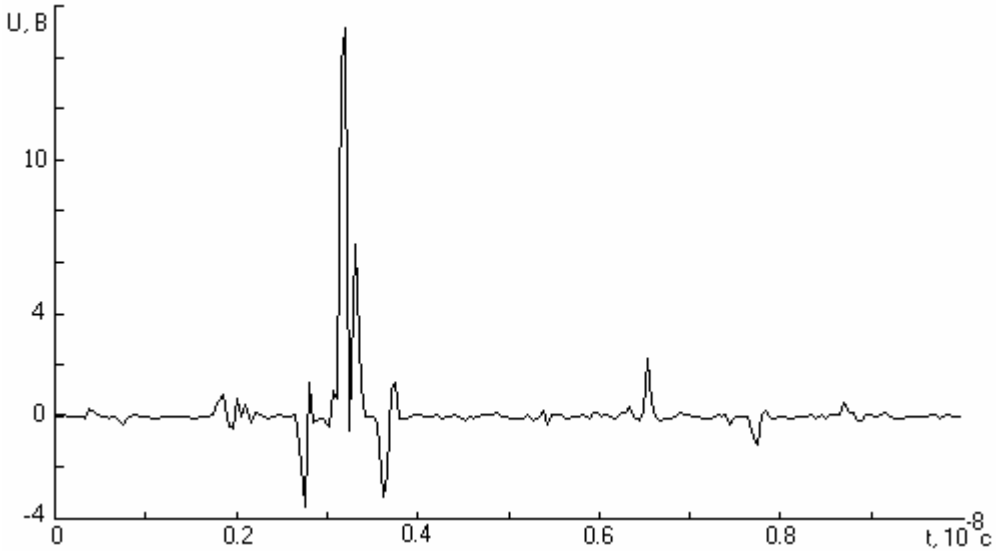


Рис. 16

Аппроксимация с оптимальным значением параметра регуляризации

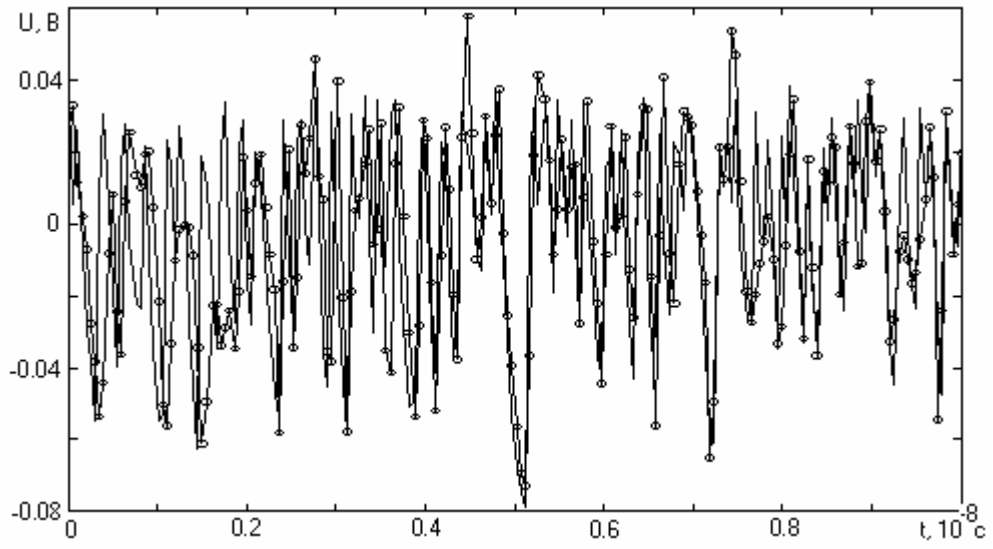


Рис. 17

Зависимость ошибок аппроксимации (e) от параметра регуляризации

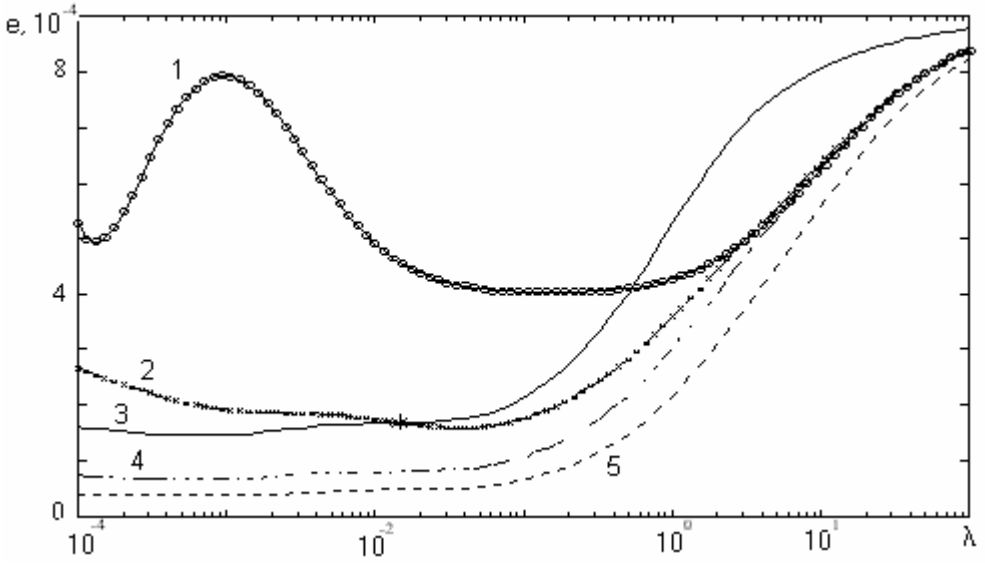


Рис. 18

Аппроксимация с оптимальным значением параметра регуляризации

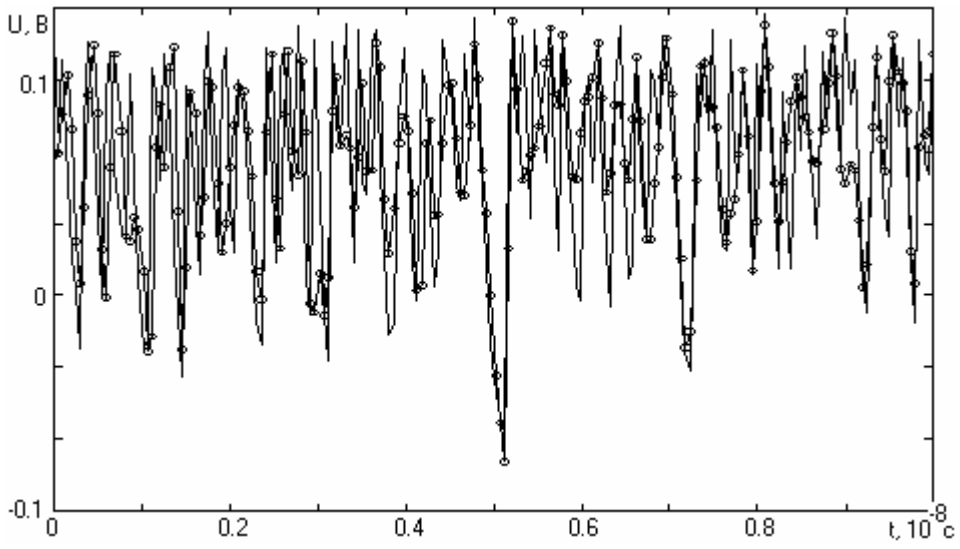


Рис. 19

Зависимость ошибок аппроксимации (e) от параметра регуляризации (λ)

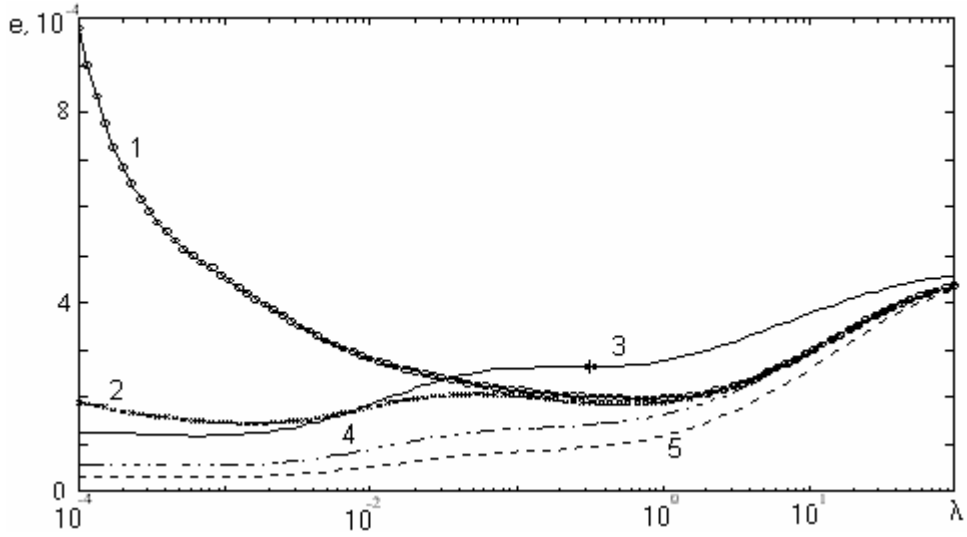


Рис. 20

Проведённые в работе исследования применимости РБФ-сетей для аппроксимации шумовых выходных сигналов кремниевых диодов показали их значительное преимущество перед классическими нейронными сетями по скорости обучения и точности аппроксимации. Показано также, что аналитический расчет РБФ-сети с использованием методик оптимизации обеспечивает точность, сравнимую с точностью, полученной при точной аппроксимации. Гладкость аппроксимирующей функции регулируется подбором параметра регуляризации.

Оптимизация сети с использованием различных критериев дает практически одинаковые значения параметров сети, различия в скорости оптимизации также незначительны. Показано, что с увеличением числа точек аппроксимации точность аппроксимации сохраняется.

3.4.2. Прогнозирование шумовых выходных сигналов кремниевых диодов

На рисунке 21 и 22 показан результат прогнозирования 20 точек на основе 200 точек данных. Прогнозируемые точки начинаются с $0.9*10^{-8}$ с.

Результаты прогнозирования шумовых выходных сигналов кремниевых диодов РБФ-сетями показывают целесообразность применения данных сетей для решения задач прогнозирования. Видно, что вид аппроксимирующей функции при прогнозировании сохраняется с очень высокой точностью, различие есть только в уровнях сигналов.

Прогнозирование РБФ-сетью с гауссовым ядром

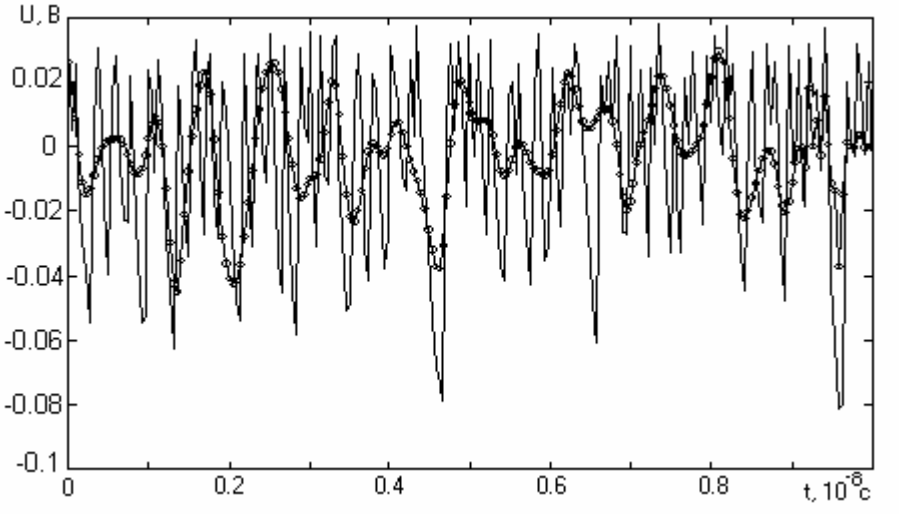


Рис. 21

Прогнозирование РБФ-сетью с мультиквадратичной функцией

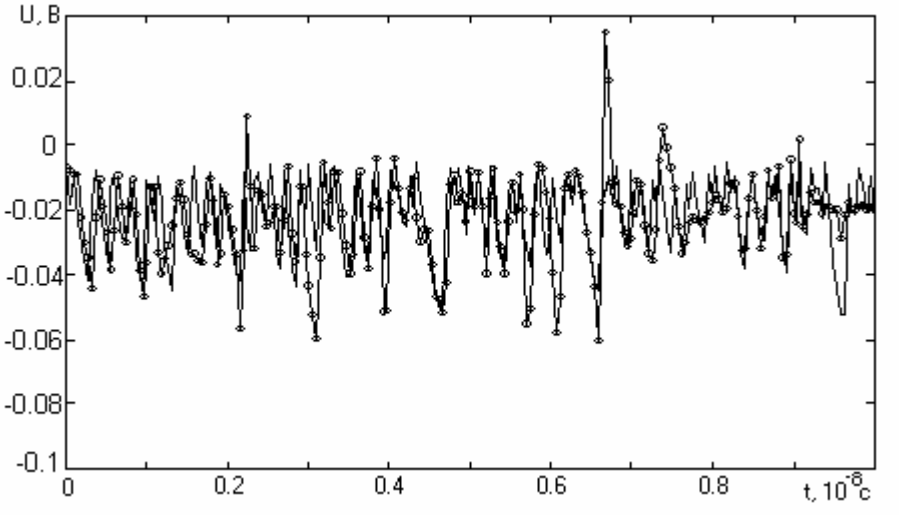


Рис. 22

3.5. Выводы

В результате проведённых исследований была показана возможность применения ИНС для решения обратной инженерной задачи, т.е. возможность восстанавливать внутренние характеристики системы по известным выходам. Кроме того, для сетей с радиальным базисом была продемонстрирована возможность аппроксимации сигналов шумовых кремниевых диодов и их предсказания на длину 10% от обучающей выборки.

4. ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ

4.1. Введение

В настоящем разделе рассмотрены практические применения нейросетевых моделей полупроводниковых приборов. Показано, что они могут быть использованы для коррекции сигналов в системах обработки информации с целью устранения аппаратных искажений. Рассмотрены два класса полупроводниковых приборов: многоэлементные фотоприемники инфракрасного диапазона и лавинные фотодиоды.

Предложены алгоритмы обработки сигналов фотоприемников, устойчивые к шумам и позволяющие повысить надежность передачи информации в оптических системах связи.

4.2. Коррекция сигналов многоэлементных фотоприёмников

Многоэлементные сенсоры применяются для регистрации оптического и инфракрасного излучения в медицине, неразрушающем контроле, экспериментальной физике и других областях науки и техники [28-30]. При их изготовлении не удается получить идеальной идентичности характеристик и исключить взаимное влияние отдельных элементов, что приводит к искажениям сигналов таких фотоприемников. Дополнительные шумы и искажения возникают в процессе коммутации чувствительных элементов [31]. Современные технологии позволяют производить однокристальные интеллектуальные сенсоры со встроенными функциями самотестирования [32]. При использовании нейросетевых алгоритмов в них возможно осуществление дополнительных функций встроенной коррекции. Модель многоэлементных инфракрасных фотоприемников диапазона 3-5 мкм, которая может быть использована для такой коррекции искажений, рассмотрена в работах [33-34]. На примере рассмотренных фотоприемных устройств показана эффективность нейросетевых методов коррекции искаженных сигналов.

4.2.1. Математическая модель

В математическом смысле выбор модели - это выбор соответствующего уравнения. Предполагая, что многоэлементный фотоприемник (МЭФП) представляет собой ансамбль или матрицу единичных фотоприемников, как показано на рисунке 23, для описания динамики его функционирования в общем случае применяют следующее векторное дифференциальное уравнение [35]:

$$dZ_d(t) = F_d[Z_d(t), t]\, dt + D_d(t)\, dN[L(t), t], \qquad (59)$$

где Z_d – m-мерный вектор состояния матрицы (каждый компонент этого вектора может рассматриваться как переменная состояния отдельного чувствительного элемента), X – m-мерный входной процесс; D_d – диагональная матрица mxm известных функций времени, F_d - - m-мерный вектор, состоящий из известных нелинейных функций, осуществляющих неинерционное преобразование вектора состояния; N - m- мерный дважды стохастический ТСП, $L(t)$ - векторная функция, определяемая интенсивностью детектируемого поля излучения.

Чувствительные элементы МЭФП подвержены различным случайным воздействиям, приводящим к флуктуациям их параметров. В зависимости от физических причин эти изменения могут носить различный характер. Если изменения этих параметров являются медленными по сравнению с временным масштабом изменения преобразуемых сигналов, то для описания МЭФП применимы уравнениями со случайными коэффициентами, принимающими случайные, но постоянные значения во всем интервале наблюдения.

Если скорость изменений сравнима со скоростью протекания процесса фотодетектирования, то МЭФП может быть описан уравнениями с коэффициентами в виде случайных функций. С помощью уравнения (59) моделировалось изменение характеристик МЭФП, состоящего из фоторезисторов и избирательных усилителей, при воздействии на него модулированного излучения. Такая задача возникла при разработке сканирующего радиометра и спектрометра инфракрасного излучения диапазона 3-5 мкм.

В нем были использованы многоэлементные фотоприемные устройства ФУЛ-123, выполненные по гибридной технологии и имеющие в своем составе 14 тонкопленочных фоторезисторов из PbSe. Для уменьшения шума и темновых токов фоторезисторы охлаждаются встроенными двухкаскадным элементами Пельтье.

Типичные характеристики этого устройства указаны в табл. 3. Индивидуальные параметры двух экземпляров ФУЛ-123 приведены в табл. 4. На основании данных, представленных в таблицах можно сделать вывод, что разброс параметров отдельных элементов для рассмотренного экземпляра (уровня сигнала U_c , и уровня шума $U_ш$,) превышает 10 % .

Представленные характеристики показывают, что проблема различий чувствительности отдельных элементов ФУЛ-123 при модуляции потока излучения не решается, а становится еще более актуальной.

Таблица 3.

Характеристики фотоприемного устройства ФУЛ-123.

Характеристика	Типичное значение
Количество чувствительных элементов	14
Размеры одного элемента, мкм	150 x 200
Спектральный диапазон чусвтвительности, мкм	3 -5
Удельная обнаружительная способность D^*(300K)	4×10^9
Уровень собственных шумов при $\Delta f = 150$, мкВ	340 - 400
Частота модуляции, Гц	400
Коэффициент внутреннего усиления	40 - 60
Энергетическая чувствительность В/Вт	100000
Мощность, рассеиваемая термоохладителем, Вт	3.5

Таблица 4.

Индивидуальные параметры ФУЛ-123 № 9008 и №31888.

Номер элемента	U_c, *мВ*		$U_ш$, *мкВ*	
	№ 9008	№ 3188	№ 9008	№ 3188
1	41	94	340	880
2	46	96	360	920
3	46	94	340	900
4	46	97	340	860
5	50	86	360	1000
6	52	92	380	940
7	51	86	380	900
8	45	96	380	880
9	42	96	380	880
10	40	90	360	820
11	38	92	340	880
12	38	85	360	860
13	53	90	380	840
14	52	94	380	900

Сравнение результатов моделирования сигналов ФУЛ-123 с использованием индивидуальных особенностей МЭФП (табл. 4) и эксперимента с эталонным источником излучения типа абсолютно черного тела показали, что в последнем случае искажения проявляются в большей степени. Это объясняется взаимными влиянием элементов, в результате которого возникает дополнительная нестационарная

составляющая шума, индуцированного сигналами соседних элементов. Построение более точных моделей МЭФП возможно с применением технологии искусственных нейронных сетей.

4.2.2. Нейросетевая модель МЭФП

Следует отметить, что учет взаимного влияния элементов значительно усложняет построение точной модели и анализ характеристик МЭФП. Кроме того, полная информации о величине коэффициента фотоэлектрической связи как правило нет в справочниках и паспортах. Этот параметр определяется как отношение значения напряжения или тока фотосигнала неосвещенного фоточувствительного элемента, расположенного рядом с освещенным фоточувствительным элементом, к значению напряжения или тока фотосигнала последнего.

Для построения модели использовалась одна из разновидностей ИНС — однослойный персептрон (ОСП). В этом случае структура и начальные значения весовых коэффициентов исходной модели выбираются на основании типичных для данного фотоприемника характеристик. При этом процедура настройки весов для индивидуальных экземпляров фотоприемника входит в алгоритм обработки данных системы, в которой используется этот конкретный фотоприемник, с его индивидуальными особенностями.

Разработанная модель позволяет оценить характеристики многоэлементных фотоприемников на этапе их проектирования. Результаты данной работы могут быть использованы при построении алгоритмов коррекции искажений, вносимых многоэлементными фотоприемниками [33-34].

4.2.3. Результаты

Описанный метод был применён к сигналам многоэлементного фотоприёмника (ФУЛ-123). Результат коррекции представлен на рисунке 25.

Оказалось, что ИНС в состоянии отследить и удалить периодические искажения и другие систематические ошибки. В то же время она не в состоянии справиться с искажениями, имеющими статистический характер.

Схема моделирования процесса формирования сигналов
в многоэлементном фотоприемнике.

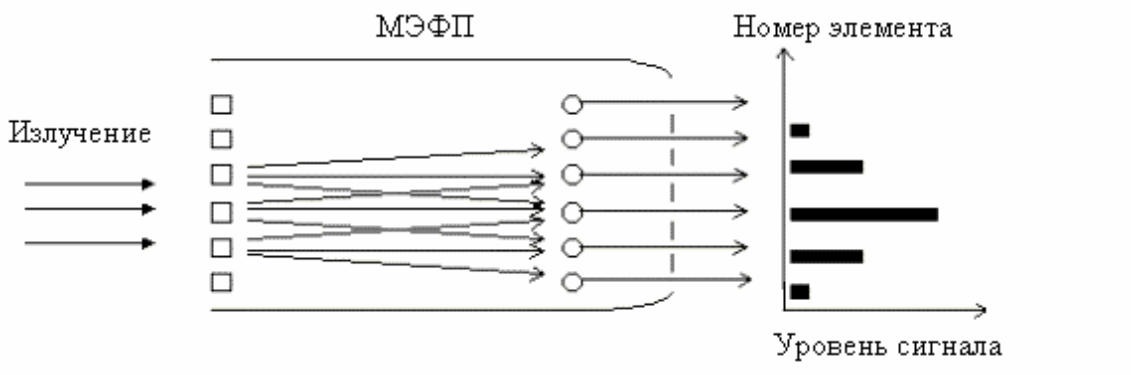


Рис. 23

Структура нейронной сети.

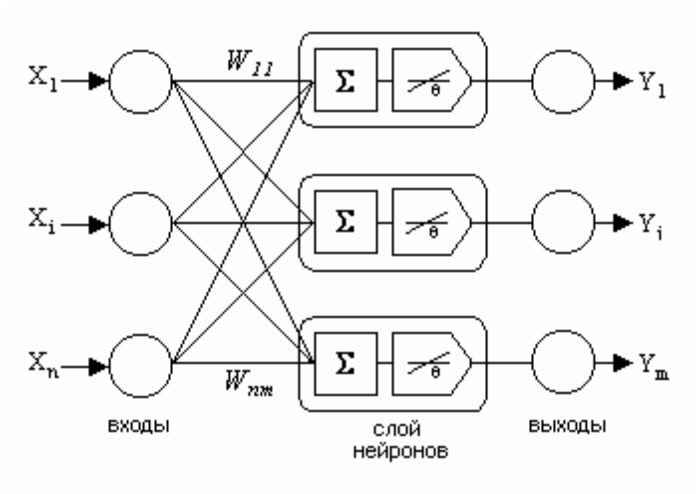
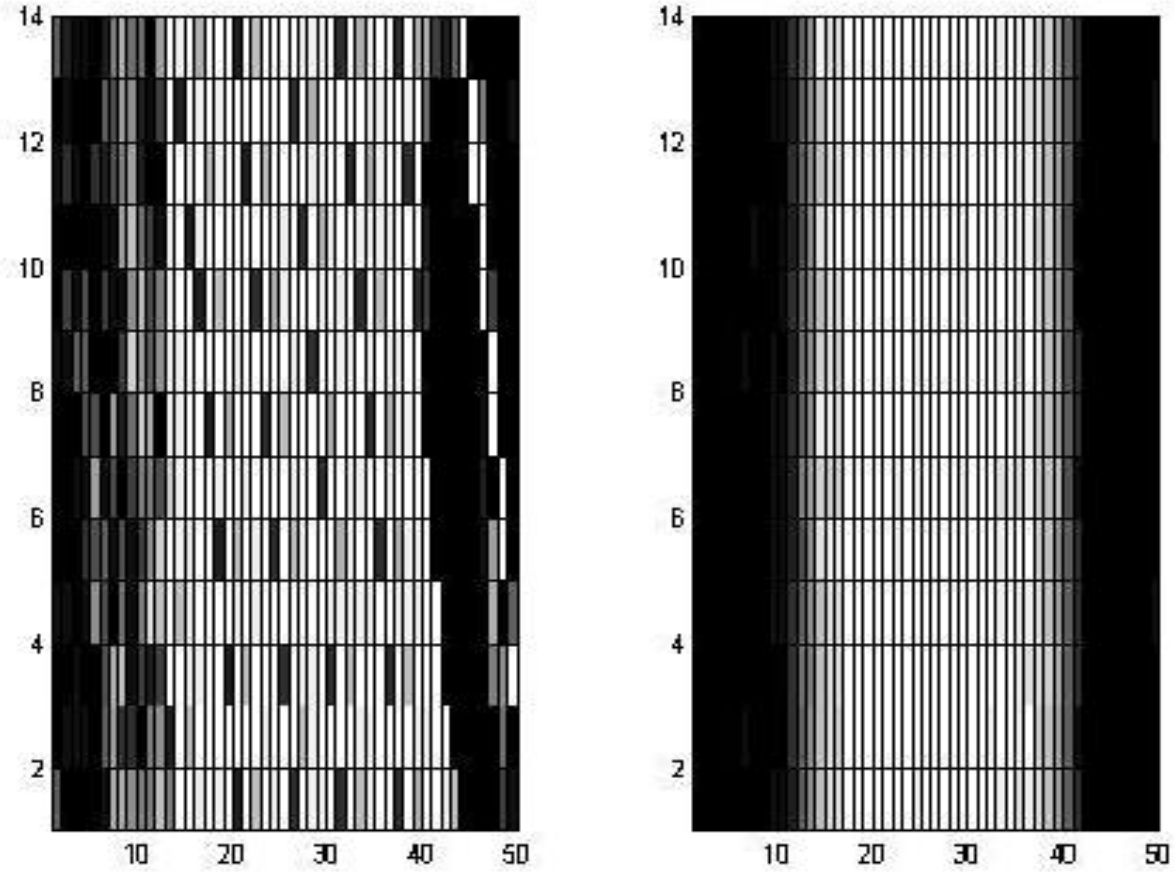


Рис. 24

Коррекция изображения жала нагретого паяльника:

а) исходное изображение, б) скорректированное изображение.



а) б)

Рис. 25

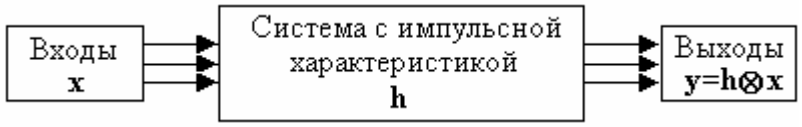
4.3. Коррекция инерционности

Нейросетевые алгоритмы обработки сигналов в технике оптической связи открывают новые перспективы при решении задач, прежде считавшихся неразрешимыми в силу сложности их формализации. Их развитие приобретает особую актуальность с появлением процессоров, оптимизированных под векторно-матричные операции [37]. В настоящей работе предлагается алгоритм восстановления интенсивности оптического излучения, регистрируемого инерционными фотодетекторами в условиях наложения однофотонных импульсов. В таком режиме известные методы обработки сигналов фотодетектора (методы интегрирования фототока и счета фотонов) приводят к недопустимо высоким ошибкам [38]. В данном случае нейронная сеть позволяет восстановить сигналы, искаженные фотодетекторами с импульсными характеристиками, отличными от δ-функции. Метод позволяет увеличить пропускную способность оптического канала передачи информации в режиме счета фотонов и может быть использован для обработки сигналов лавинных фотодиодов (ЛФД).

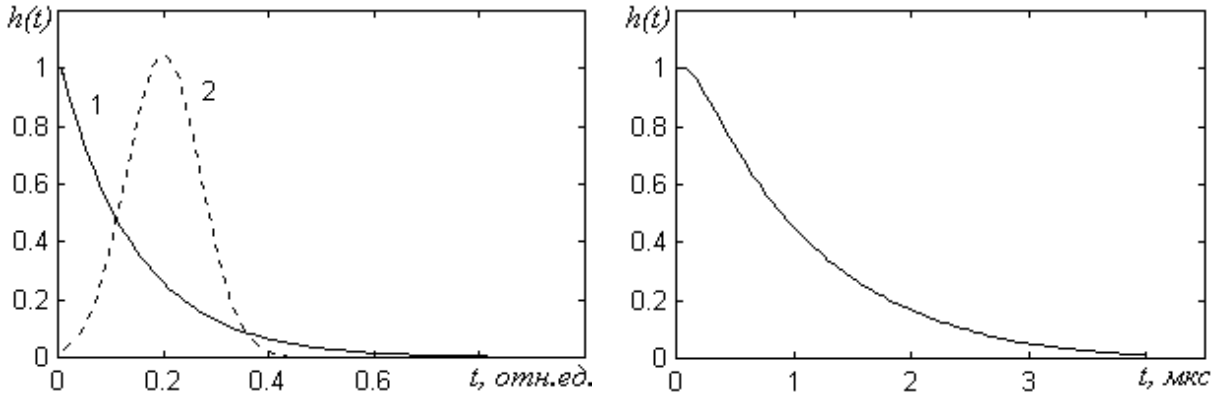
Рассматривается следующая задача. На вход фотодетектора с импульсной характеристикой $h(t)$ поступает нестационарный поток фотонов $x(t)$, полагаемый пуассоновским, см. рисунок 26. Выходной сигнал y определяется в виде: $y(t)=x(t)\otimes h(t)$, где символом $\otimes$ обозначена операция свертки.

Реакцией фотодетектора на каждый из фотонов является импульс, форма которого определяется его типом и параметрами цепи нагрузки. Спад каждого из таких однофотонных импульсов затянут и имеет форму близкую к экспоненциальной. При достаточно высокой частоте следовании импульсов они перекрываются, что в значительной степени затрудняет их разделение. Такой режим работы фотодетекторов обычно не используется из-за сложностей обработки данных и условно его можно считать переходным между режимом счёта фотонов и режимом интегрирования фототока. Подход, предлагаемый ниже, позволяет решить рассматриваемую задачу путем разделения исходных однофотонных импульсов с помощью нейронной сети.

Инерционная система (*а*), идеализированные импульсные характеристики (*б*) и характеристика реального фотодетектора (*в*)



а)



б) *в*)

Рис. 26

Двухкаскадная сеть с блоком задержки

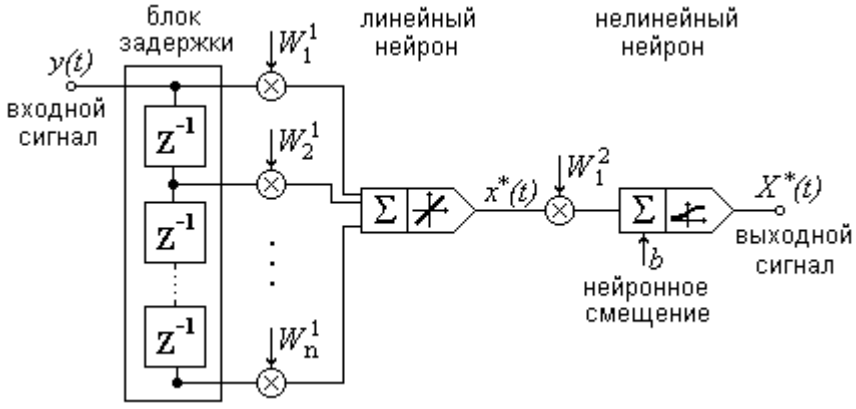


Рис.27

Для восстановления исходного сигнала на основании искажённого выходного используется двухкаскадная нейронная сеть. Ее первый каскад представляет собой адаптивный линейный нейрон, второй – нелинейный нейрон с логистической выходной функцией. Сеть изображена на рисунке 27.

Блок задержки и линейный нейрон решают задачу деконволюции подаваемого сигнала $y(t)$. Эта часть сети аналогична цифровому фильтру с конечной импульсной характеристикой и настраиваемыми параметрами W_i^1. Результатом деконволюции является скорректированный сигнал $x^*(t)$.

На практике сигнал $x^*(t)$ определяется с ошибкой в силу наличия в выходном сигнале $y(t)$ фоновых импульсов и шумов. Поэтому необходимо выбирать оптимальный уровень дискриминации, ниже которого импульсы в сигнале $x^*(t)$ считаются шумовыми, а выше – однофотонными лавинами. В качестве элемента, производящего это разделение был выбран нелинейный нейрон. Весовой коэффициент W_1^2 влияет на чёткость границы раздела, смещение b – на её положение. На рисунке 27 сигналу, в котором произведено разделение импульсов, соответствует выходной сигнал $X^*(t)$. В общем случае вместо рассмотренных каскадов, в каждом из которых содержится по одному нейрону, и блока задержки можно использовать каскады с большим числом нейронов, что позволяет ускорить работу сети, однако усложняет ее аппаратную реализацию.

При обучении сети вначале независимо настраивались веса линейного нейрона, а затем – весовой коэффициент и нейронное смещение во втором каскаде. Процесс расчёта весов для первого каскада в общем случае производится следующим образом. Пусть линейная сеть, тренируемая на l обучающих парах, имеет n входов и m выходов. Матричное уравнение преобразования сигналов для такой сети имеет вид:

$$T_{m,l} = W_{m,n} P_{n,l} + b_{m,1} e_{1,m}, \tag{60}$$

где $P_{n,l}$, $T_{m,l}$ – обучающие пары, $W_{m,n}$ – матрица весов, $b_{m,1}$ – вектор смещений нейронного слоя, $e_{1,m}$ – вектор единиц.

Если ввести замены

$$A_{n+1,l} = \begin{bmatrix} P_{n,l} \\ 1 \end{bmatrix}, \tag{61}$$

$$X_{m,n+1} = \left[W_{m,n} \quad b_{m,1} \right], \tag{62}$$

то выражение (60) примет вид

$$T_{m,l} = X_{m,n+1} A_{n+1,l} \ . \tag{63}$$

Тогда матрица $X_{m,n+1}$ может быть выражена

$$X_{m,n+1} = T_{m,l} \, A^{+}_{l,n+1}, \tag{64}$$

где $A^{+}_{l,n+1}$ – псевдообратная матрица по отношению к матрице (61). Теперь искомые матрицы $W_{m,n}$ и $b_{m,1}$ запишутся в виде

$$W_{m,n} = \begin{bmatrix} x^{1,1} & \cdots & x^{1,n} \\ \vdots & & \vdots \\ x^{m,1} & \cdots & x^{m,n} \end{bmatrix}, \qquad b_{m,1} = \begin{bmatrix} x^{1,n+1} \\ \vdots \\ x^{m,n+1} \end{bmatrix}, \tag{65}$$

где $x^{i,j}$ – элементы матрицы $X_{m,n+1}$.

Важным моментом при обучении оказался оптимальный выбор сигналов, по которым обучается сеть (т.н. обучающих пар).

Работа сети тестировалась на сигнале длинной в 25600 отсчётов, что соответствует 400 парам. Пример восстановления искусственно искажённого сигнала представлен на рисунке 28.

Далее было проведено исследование устойчивости предлагаемого метода обработки сигналов к шумам. Для этого искаженный сигнал $y(t)$ суммировался с гауссовым шумом, и в таком виде поступал на вход ИНС. Опыт показал, что метод работает вплоть до отношения сигнал/шум (ОСШ) порядка 50, что (в данном опыте) соответствует гауссову шуму с среднеквадратичным отклонением шума $\sigma = 0.02$. На рисунке 29 представлена зависимость относительного количества ошибочно определённых пиков от среднеквадратичного отклонения (σ) шума.

На рисунке 30 показан результат восстановления реального сигнала, полученного с помощью лавинного фотодиода ФД-115Л.

Восстановление искусственного сигнала при
отношении сигнал/шум, равном 100.

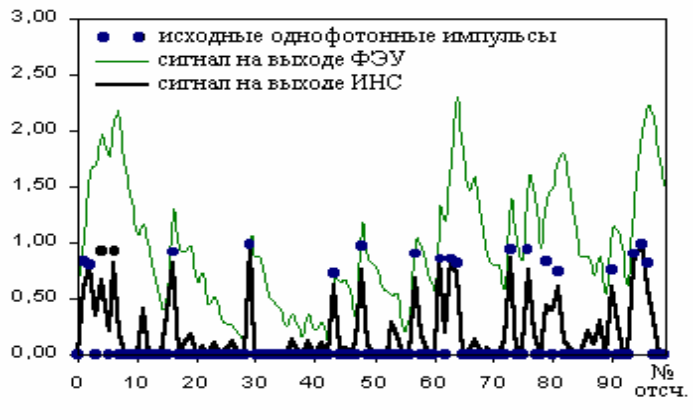


Рис. 28

Зависимость относительного количества ошибок от σ шума
(1- обучающие сигналы, 2 – тестовые сигналы).

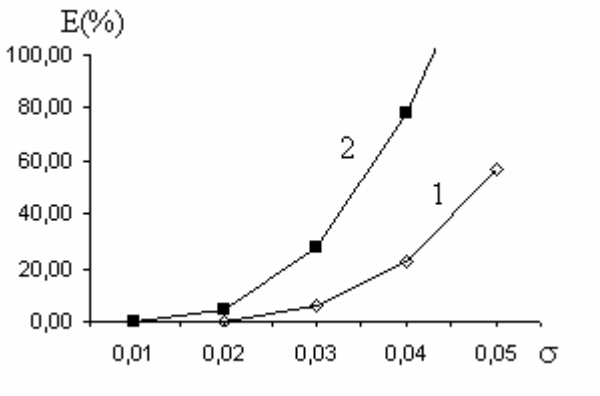
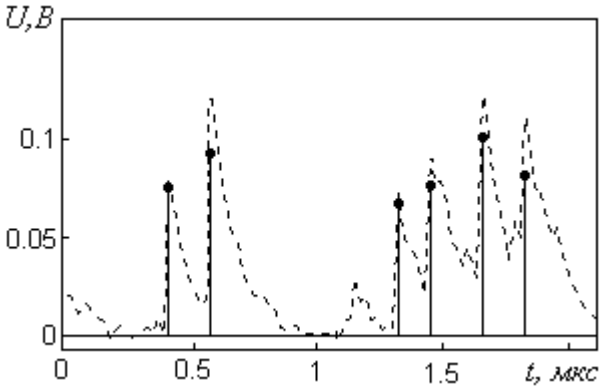


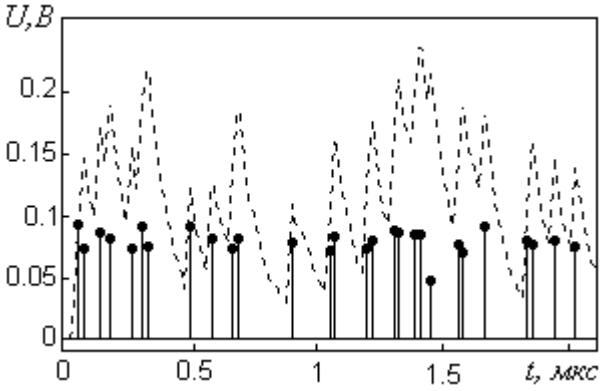
Рис. 29

Восстановление сигнала ФД 115Л:

а) обучающий сигнал, б) тестовый сигнал.



а)



б)

Рис. 30

Необходимо отметить, что обучение сети проводилось на сигнале, для которого справедлив метод счёта фотонов. Рассмотренный алгоритм обработки сигналов фотоприемников устойчив к шумам вплоть до ОСШ порядка 100-50, что позволяет повысить надежность передачи информации в оптических системах связи.

4.4. Выводы

В этом разделе было показано, что нейронные сети способны восстанавливать сигналы, искажённые детерминированными операторами регистрирующих систем: взаимным влиянием элементов и инерционными импульсными характеристиками. Свойствами сетей при этом являются достаточная устойчивость к шумам и негромоздкость преобразований сигнала, что позволяет говорить о возможности их аппаратной реализации.

Кроме того, методика ИНС может быть полезна при анализе скрытых связей и влияний внутри МЭФП и при анализе однофотонных лавин лавинных фотодетекторов.

ЗАКЛЮЧЕНИЕ

В настоящей работе разработан новый непараметрический подход к построению моделей процессов, протекающих в полупроводниковых приборах, этот подход основан на применении ИНС. Основные результаты НИР заключаются в следующем:

1. Проведен анализ традиционных подходов в моделировании полупроводниковых приборов. Рассмотрены преимущества математического аппарата нейронных сетей в моделировании и анализе полупроводниковых приборов.
2. Экспериментально исследованы флуктуационные характеристики выходных сигналов диодов кремниевых генераторных в различных режимах их работы.
3. Построены модели на основе нейронных сетей для аппроксимации и прогнозирования выходных сигналов полупроводниковых диодов. Проведен анализ эффективности этих моделей на основе машинного эксперимента с использованием данных физического эксперимента.
4. Построена нейронная сеть для определения режимов работы и характеристик полупроводниковых диодов. Данная сеть протестирована на основе экспериментальных данных.
5. Разработаны нейросетевые модели для коррекции сигналов в системах обработки информации с целью устранения аппаратных искажений. Построена модель многоэлементного фотоприемника с возможностью устранения систематических ошибок. Разработана модель для коррекции инерционности фотодетекторов. Предложенные модели использованы в коррекции данных физического эксперимента и показали хорошие результаты.

Разработанный и апробированный нейросетевой подход позволяет получать импульсные характеристики полупроводниковых приборов, которые могут быть использованы в задачах идентификации, оценки и прогнозирования параметров работы и других характеристик приборов. Результаты тестов показали, что разработанные модели и методы обработки сигналов приборов могут с успехом применяться в системах параллельных вычислений или нейрокомпьютерах, что резко увеличит скорость обработки данных.

Таким образом, главные задачи, сформулированные в задании на научно-исследовательскую работу, выполненную в рамках настоящего гранта, решены. Более того, намечены перспективы по применению и усовершенствованию разработанных методов и моделей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Сугано Т., Икома Т., Такэнси Е. Введение в микроэлектронику: Пер с яп. - М.: Мир, 1988. - 320 с.
2. Мулярчик С.Г. Численное моделирование микроэлектронных структур. – Минск: Изд–во Университетское, 1987. – 368 с.
3. Борздов В.М., Комаров Ф.Ф. Моделирование электрофизических свойств твердотельных слоистых структур интегральной электроники.– Мн.: БГУ, 1999.– 235 с.
4. Теория нейронных сетей. Кн.1 Учебное пособие для вузов / Ред. А.И.Галушкин. - М.: ИПРЖР, 2000. - 416 с.
5. Уоссерман Ф. Нейрокомпьютерная техника: теория и практика. – М.: Мир, 1992.
6. Гулд Х., Тобочник Я. Компьютерное моделирование в физике: В 2-х частях. М.: Мир, 1990.
7. Бусленко Н.П. Моделирование сложных систем. М.: Наука, 1978. 400 с.
8. Апанасович В.В, Тихоненко О.М. Цифровое моделирование стохастических систем. Мн.: Университетское, 1986. 127 с.
9. Шеннон Р. Имитационное моделирование систем. М.: Мир, 1978.
10. Orr M. Introduction to Radial Basis Function Networks.- University of Edinburg. –1995.
11. Советов Б.Я., Яковлев С.А. Моделирование систем: Учебник для вузов по спец. «Автоматизированные системы управления». – М.: Высшая школа, 1985. – 271 с.
12. Грехов И.В. , Сережкин Ю.Н. Лавинный пробой p-n – перехода в полупроводниках – Л.: Энергия, 1980.
13. Гафийчук В.В., Дацко Б.И., Кернер Б.С., Осипов В.В. // Физика и техника полупроводников. – 1990, № 24. – С. 724.
14. Дацко Б. И. // Физика и техника полупроводников. – 1997, № 31. – С. 186.
15. Барановский О.К., Кучинский П.В., Лутковский В.М., Петрунин А.П. Влияние облучения на кинетику микроплазм в кремниевых диодах. // Труды X Межнационального совещания «Радиационная физика твердого тела». Севастополь, 3-8 июля 2000. С. 455 – 459.
16. Барановский О.К. Особенности моделирования процесса лавинного умножения носителей заряда в микроплазмах. – Физика конденсированных сред // Тезисы докладов VIII Республиканской научной конференции студентов и аспирантов / Под ред. В.А.Лиопо. – Гродно: ГрГу, 2000. С. 20.

17. Brennan K.F., Park D.H., Wang Y.. Design and comparison of advanced semiconductor devices using computer experiments: application to APD's and HEMT's //IEEE Trans. on Elactron Dev. 1990, V.37, No 3, 536-547.
18. Биндер К. Методы Монте-Карло в статистической физике: Пер. с англ. – М.: Мир, 1982. – 400 с.
19. Пешель М. . Моделирование сигналов и систем: Пер. с англ. – М.: Мир, 1981. – 304с.
20. Zadeh L. A. // Fuzzy Sets, Information and Control. – 1965, № 8. P. 228.
21. Батороев К.Б. "Кибернетика и метод аналогий" М.: Высшая школа, 1974 год, с. 200.
22. Головко В.А. Нейроинтеллект: теория и применение. Брест, Изд. БПИ, 1999.
23. Змитрович А.И. Интеллектуальные информационные системы. – Мн.: НТООО "ТетраСистемс", 1997.
24. Bishop M. Neural Networks for Pattern Recognition, — Oxford: Clarendon Press, 1997.
25. Mulgrew Bernard Applying Radial Basis Functions // IEEE Signal Processing Magazine, 1996, № 3.
26. S. Chen, C.F.N. Cowan, and P.M. Grant. Orthogonal least squares learning for radial basis function networks. //IEEE Transactions on Neural Networks, 2(2):302-309, 1991.
27. Полупроводниковые приборы: Диоды, тиристоры, оптоэлектронные приборы: Справочник / Под ред. Н.Н. Горюнова.– М.: Энергоатомиздат, 1987.
28. Свиташев К.К., Чикичев С.И. Полупроводниковые многоэлементные фотоприемные устройства для тепловидения // Автометрия, 1998. № 4, с.с. 3-4.
29. Медицинский тепловизор на основе матричного ФПУ128х128, работающий в диапазоне спектра 2,8 - 3,05 мкм/ Курышев Г.Л., Ковчанцев А.П., Вайнер Б.Г., Гузев А.А. и др.// Автометрия, 1998. № 4, с.с. 5-12.
30. Портативный быстродействующий тепловизор на основе фокальной матрицы МДП-конденсаторов на InAs / Курышев Г.Л., Ковчанцев А.П., Вайнер Б.Г. , Гузев А.А. и др.// Автометрия, 1998. № 4, с.с. 3-4.
31. High sensitivity readout of 2D a-Si image sensors/ Fujieda I., Street A., Weisfield R.L., Nelson S., Nylen P., Perez-Mendez V., Cho G. // Jpn. J. Appl. Phys., 1993. V. 32, part 1, № 1A, p.p. 198-204.
32. Travis B. Smart sensors. - EDN, 1996, № 5, p.p. 57-60, 64-65.
33. Apanasovich V., Gordeev A., Lashina T., Lootkovski V. Computer Simulation of Image Distortions Caused by Detector Array Defects. // Pattern Recognition and Information Processing (PRIP-97).20-22 May 1997. Minsk. - Vol. 1.- P.205-208.

34. Lutkovski V., Lashina T., Gordeev A. Correction of Image apparatus distortions. // Proceedings of 2nd International Conference on Computer Methods and Inverse Problems in Non-Destructive Testing and Diagnostics. Minsk, 1998, p 567-570.
35. Солодов А.В., Солодов А.А. Статистическая динамика систем с точечными процессами. – М.: Наука, 1988. – 256 с.
36. Geman S., Bienenstock E., Doursat R.. Neural networks and the bias/variance dilemma.// Neural Computation, 1992, №4(1). P.1-58.
37. Гальярди Р.М., Карп Ш. Оптическая связь: Пер. с англ./ Под ред. А.Г. Шереметьева.– М.: Связь, 1978. – 424 с.
38. Одноэлектронные фотоприемники / С.С. Ветохин, И.Р. Гулаков, А.Н. Перцев, И.В. Резников – М.: Атомиздат, 1979.