Нейросетевой распознаватель фонем на многопроцессорной графической плате

Бондаренко И.Ю.,

Федяев О.И.

Донецкий национальный технический университет

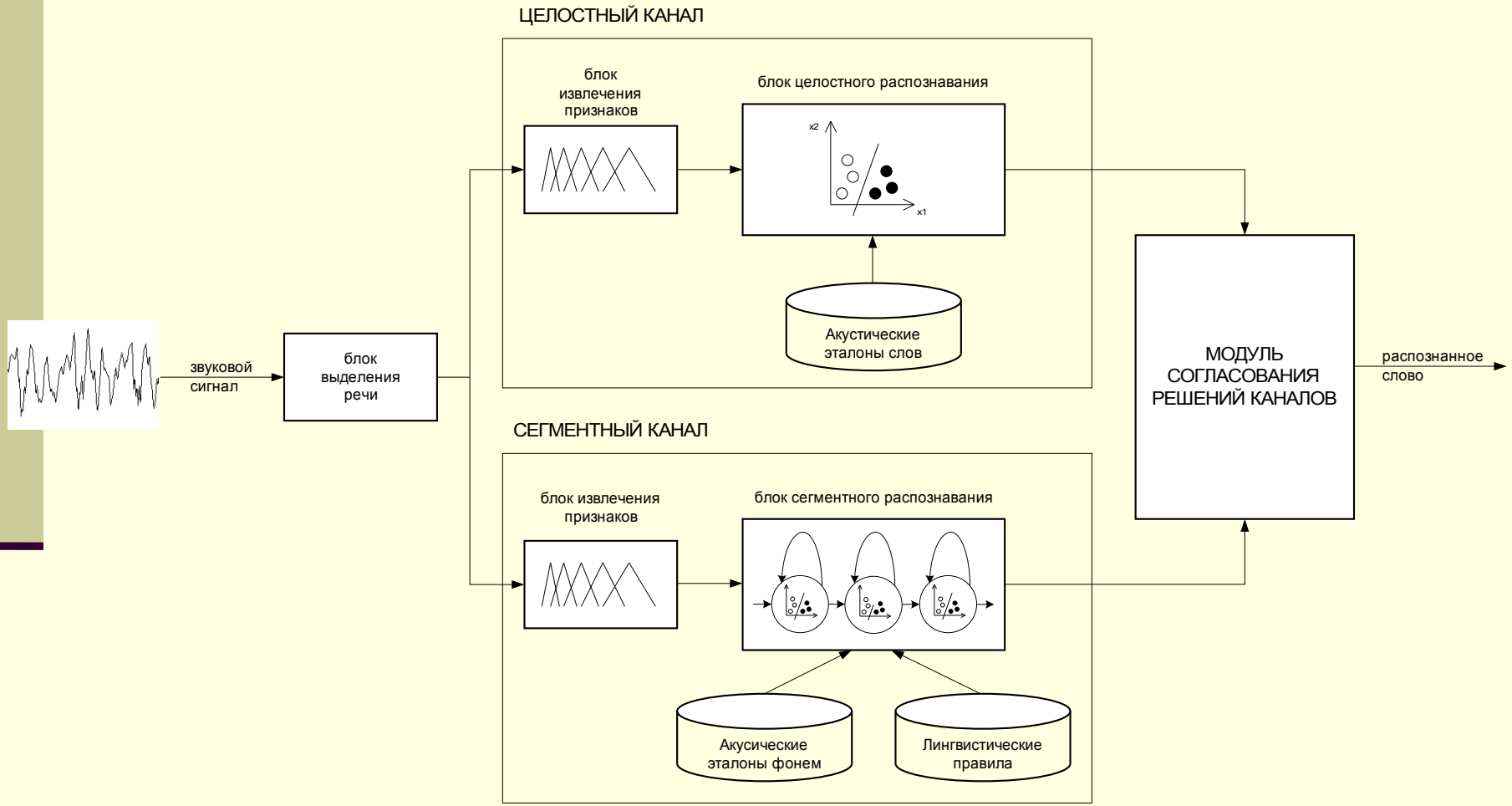
Цель работы

Целью работы является разработка алгоритмов и программно-аппаратных средств распознавания устной речи, работающих максимально быстро (в реальном масштабе времени).

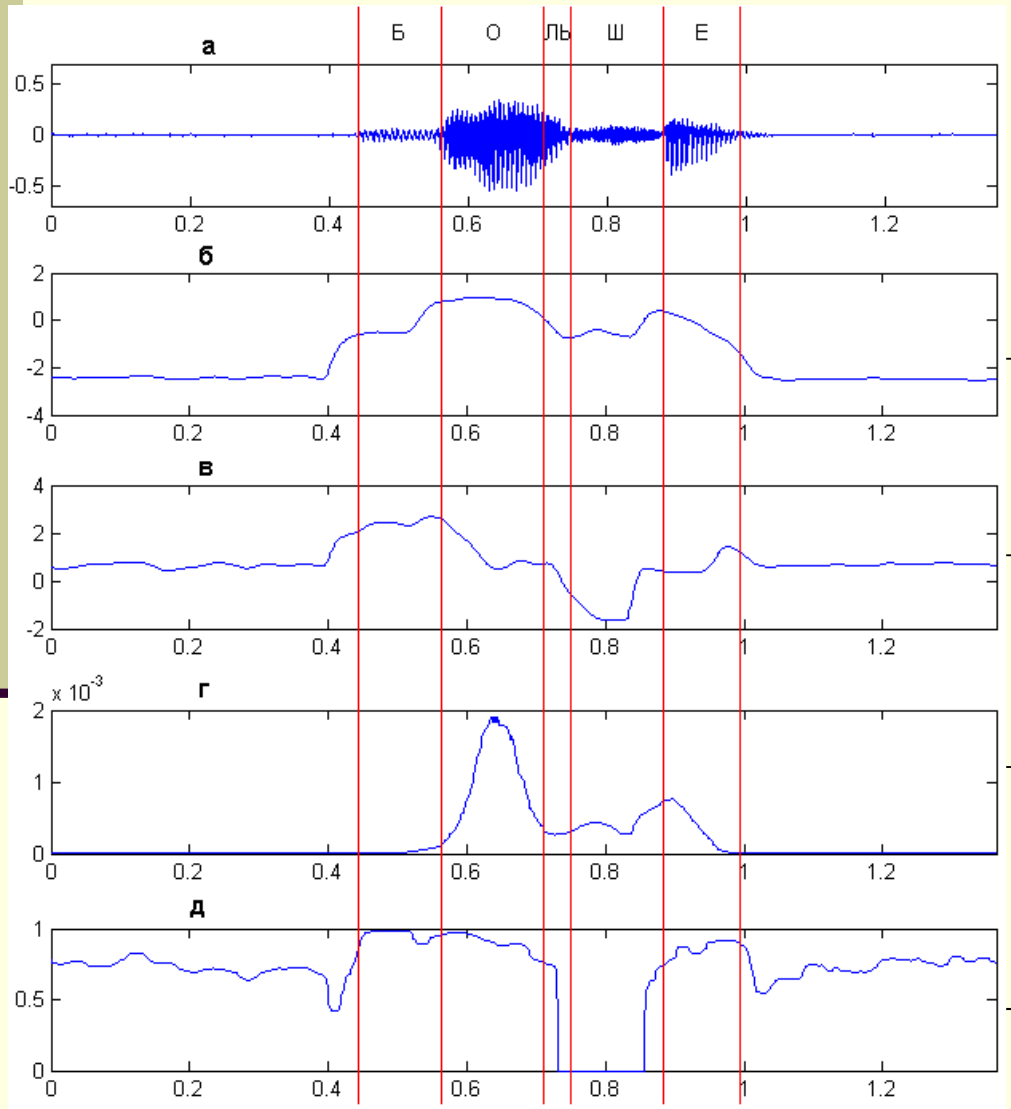
Задачи работы

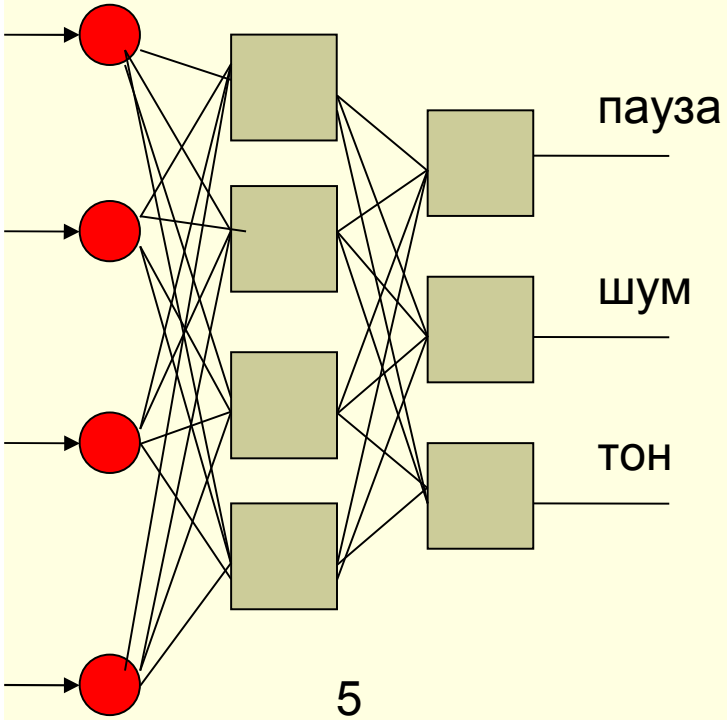
- Разработать алгоритм структурного (пофонемного) распознавания речевых слов в нейросетевом базисе
- Ускорить работу нейросетевого детектора фонем за счёт распараллеливания процессов обучения и распознавания для различных параметров нейроалгоритма

Структура сегментно-целостной системы коллективного распознавания речи

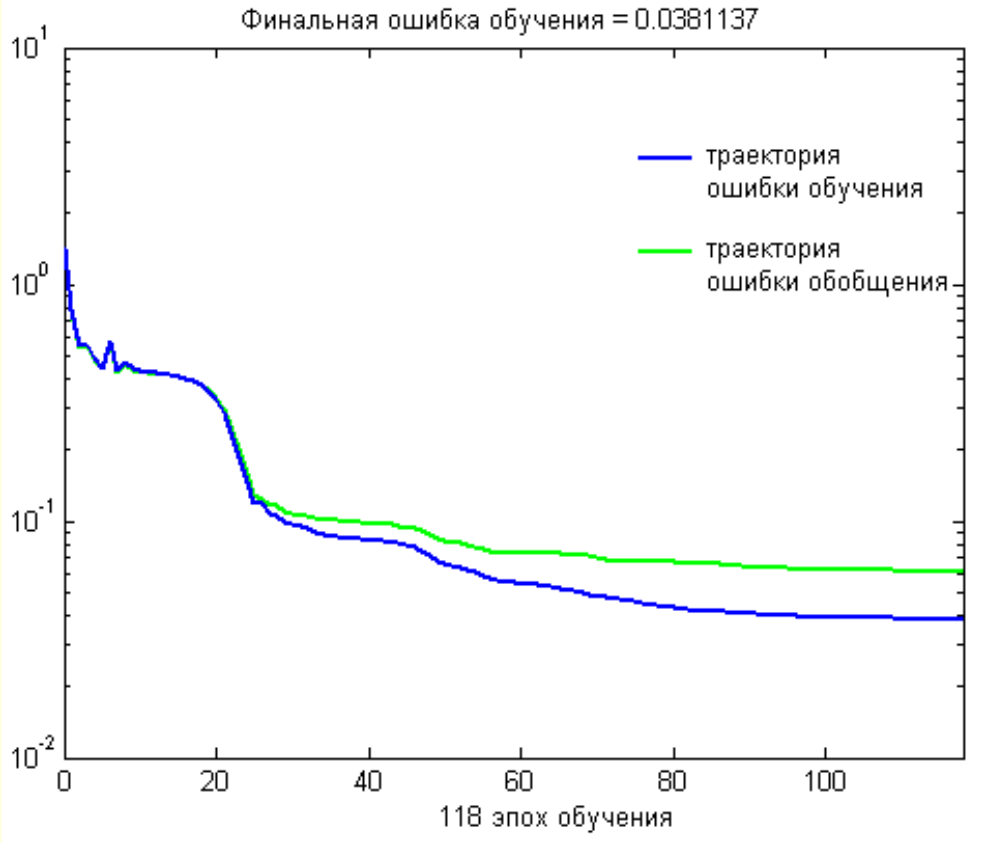


Блок выделения речевых фрагментов в звуковом сигнале





Результаты работы нейросетевого детектора «тон/шум/пауза»

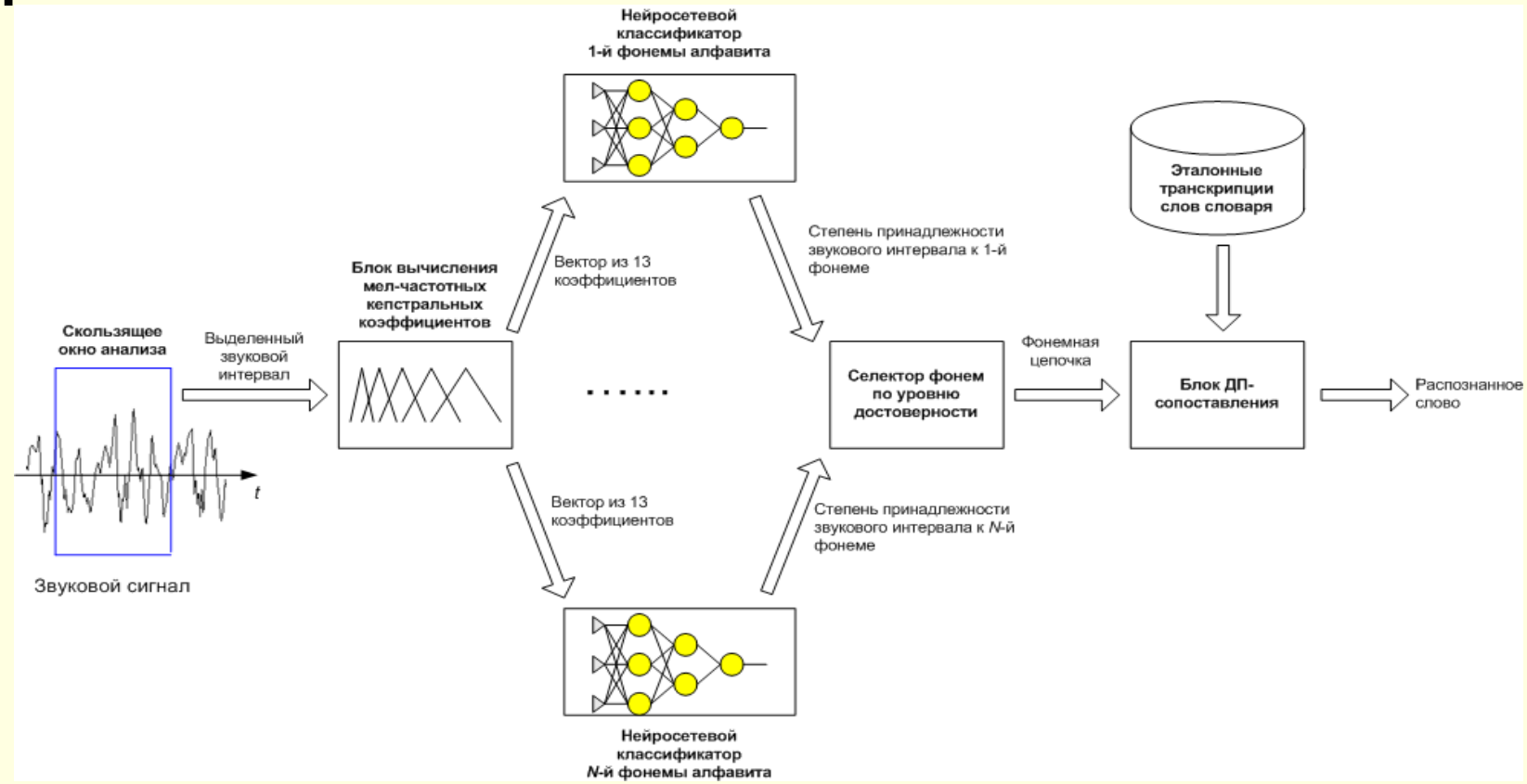


На материале речевых сигналов, записанных тремя разными дикторами с частотой дискретизации 11025 Гц в формате WAV PCM, было составлено обучающее, проверочное и тестовое множества.

Точность классификации звукового сигнала на паузные, невокализованные и вокализованные участки, определённая на тестовом множестве примеров, составила **95,49%**.

Структура сегментного канала распознавания

Основная идея: фонетический анализ речевого сигнала, основанный на **методе нейросетевого детектирования фонем**



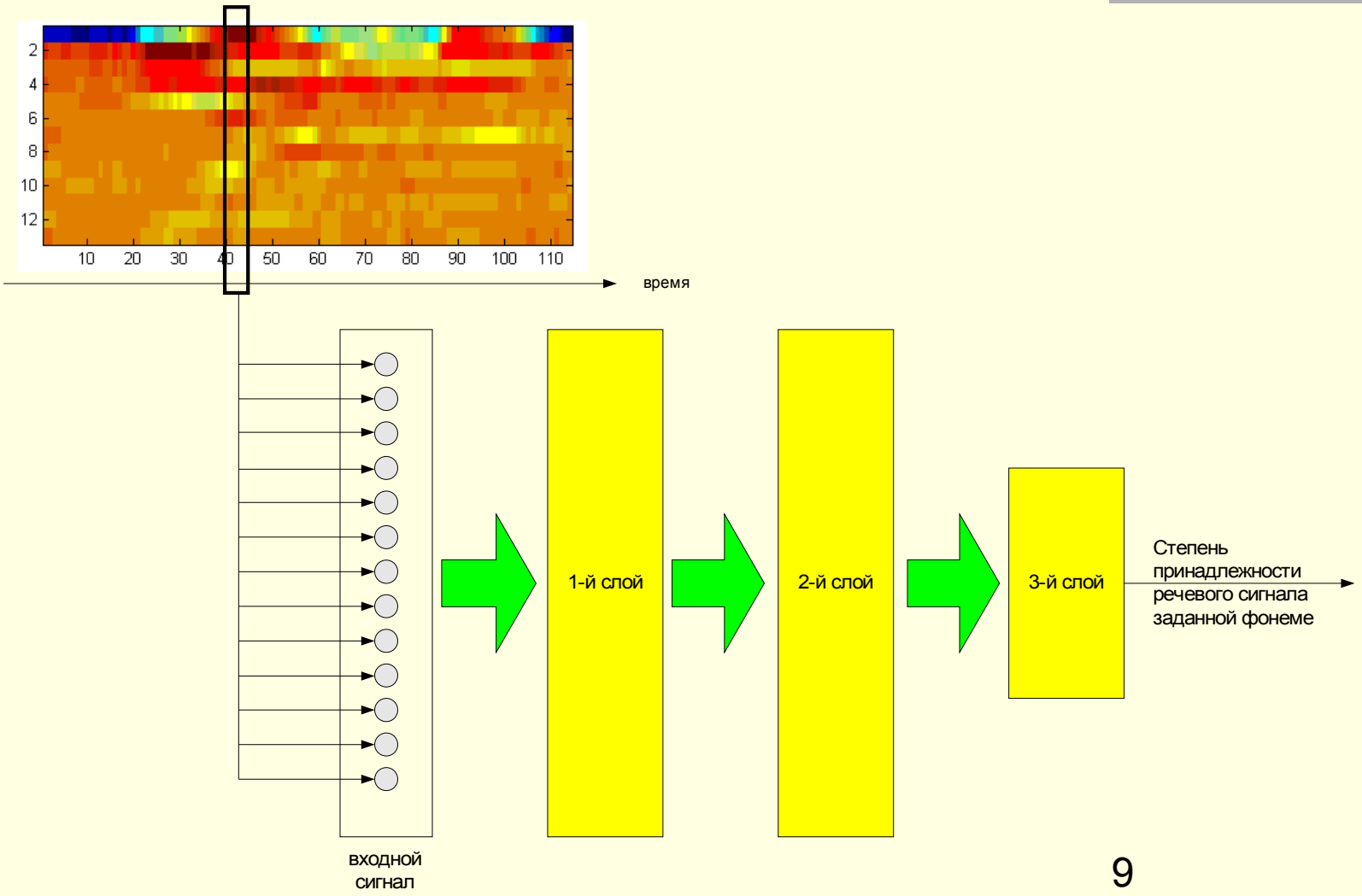
Пошаговый алгоритм нейросетевого детектирования фонем

Шаг 1. Пока окно длиной 25 мсек, скользящее вдоль оси времени шагом 5 мсек, не достигло конца речевого сигнала, выполняется мел-частотный кепстральный анализ фрагмента сигнала, вырезаемого данным окном.

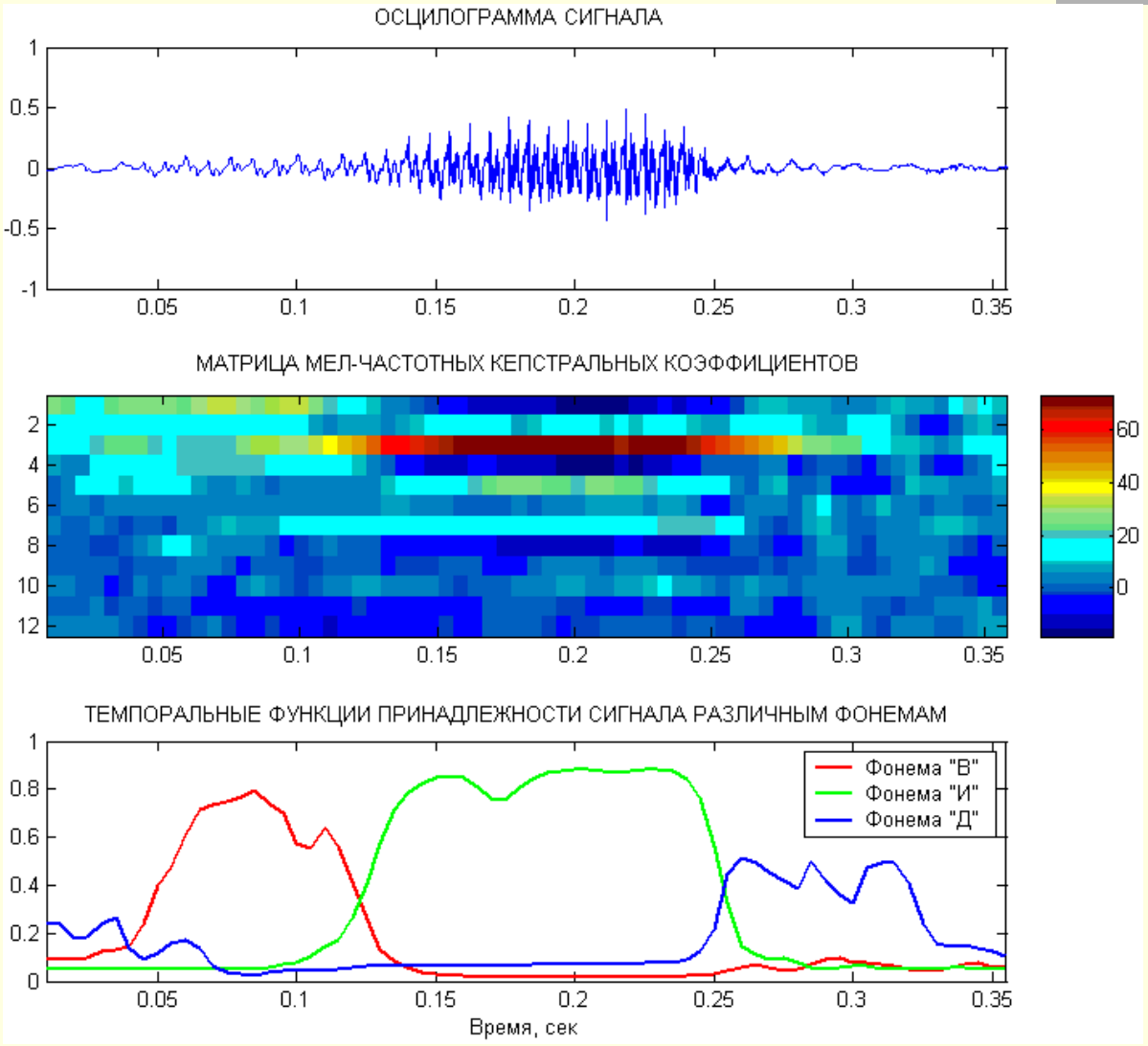
Шаг 2. С помощью нейронных сетей одновременно вычисляются степени принадлежности анализируемого звукового фрагмента ко всем фонемам фонемного словаря. Каждая нейросеть реализует функцию принадлежности для одной фонемы. Таким образом, общее число нейронных сетей – детекторов фонем составляет 42 (по числу фонем русского языка).

Шаг 3. В полученном на предыдущем шаге векторе степеней принадлежности определяется максимум. Местоположение максимума кодируется единицей, остальные элементы вектора степеней сходства кодируются нулями. Выполняется переход на шаг 1.

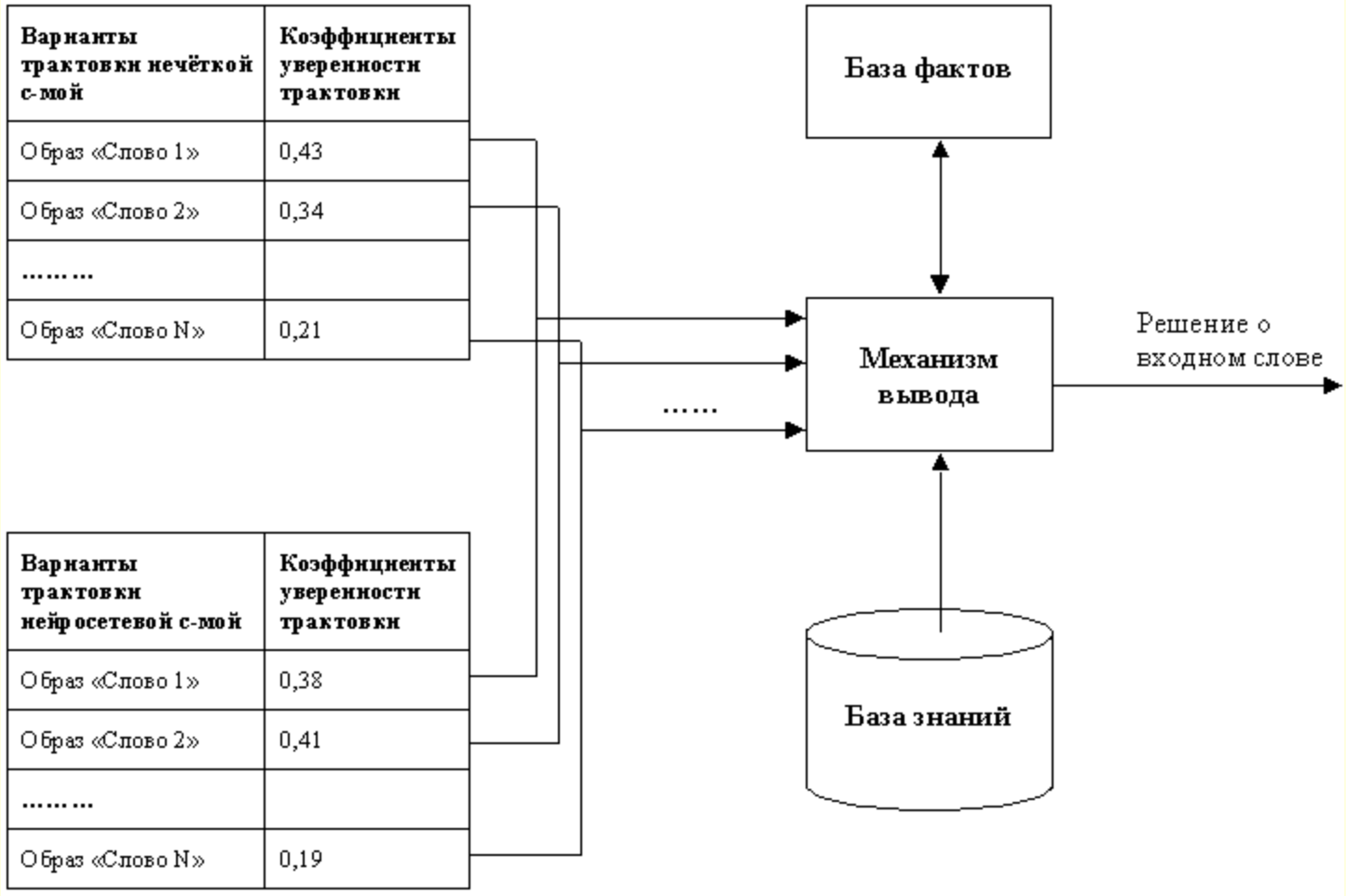
Структура нейросетевого детектора фонемы



Темпоральные функции принадлежности звукового сигнала слова «Вид» к фонемам «В», «И», «Д»



Экспертное заключение по решениям каналов распознавания



База знаний

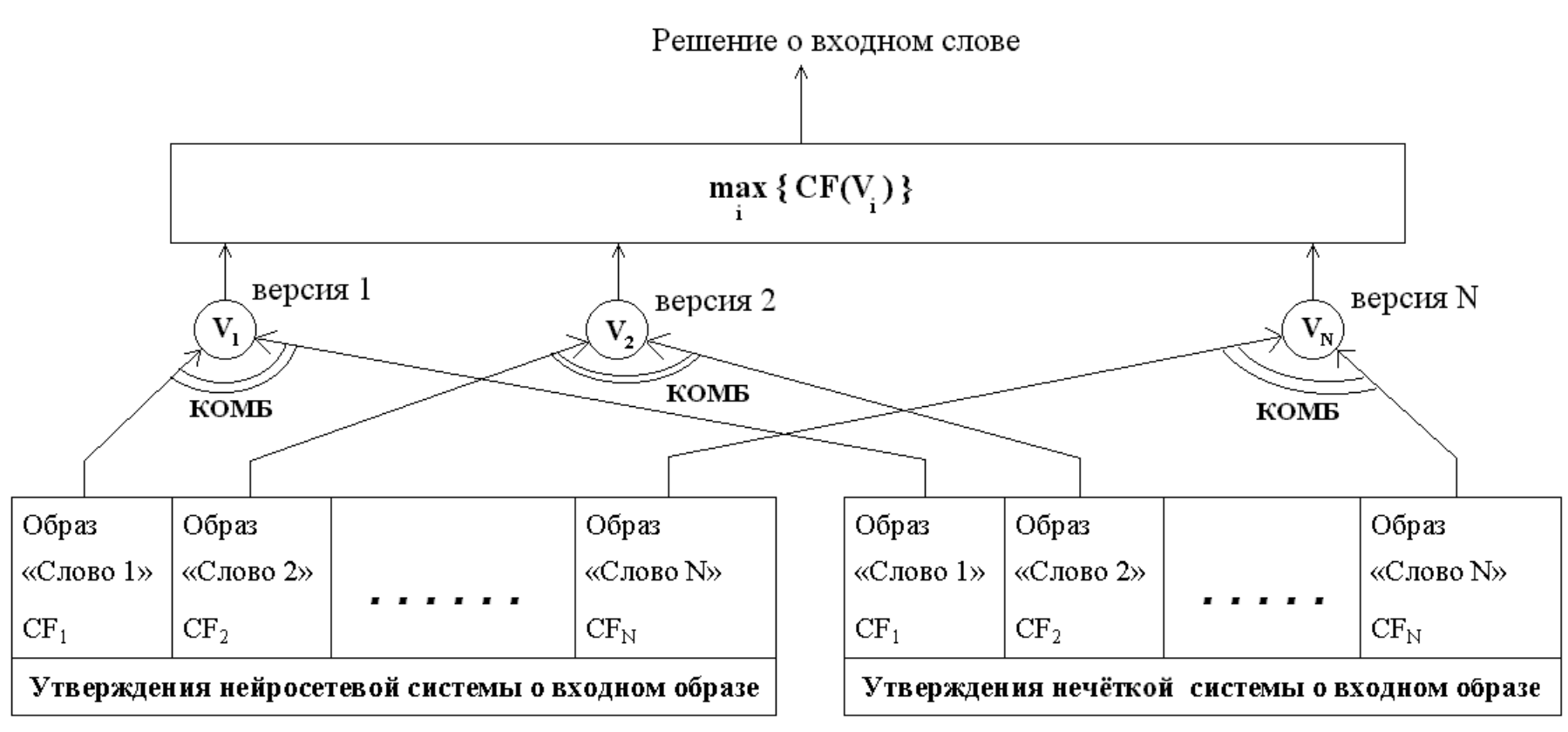
1. ЕСЛИ Трактовка нейросетевой системой входного образа = образ «Слово 1»
   ТО Версия 1 = «Слово 1»
2. ЕСЛИ Трактовка нейросетевой системой входного образа = образ «Слово 2»
   ТО Версия 2 = «Слово 2»

. . . . . .

8. ЕСЛИ Трактовка нечёткой системой входного образа = образ «Слово 1»
   ТО Версия 1 = «Слово 1»
9. ЕСЛИ Трактовка нечёткой системой входного образа = образ «Слово 2»
   ТО Версия 2 = «Слово 2»

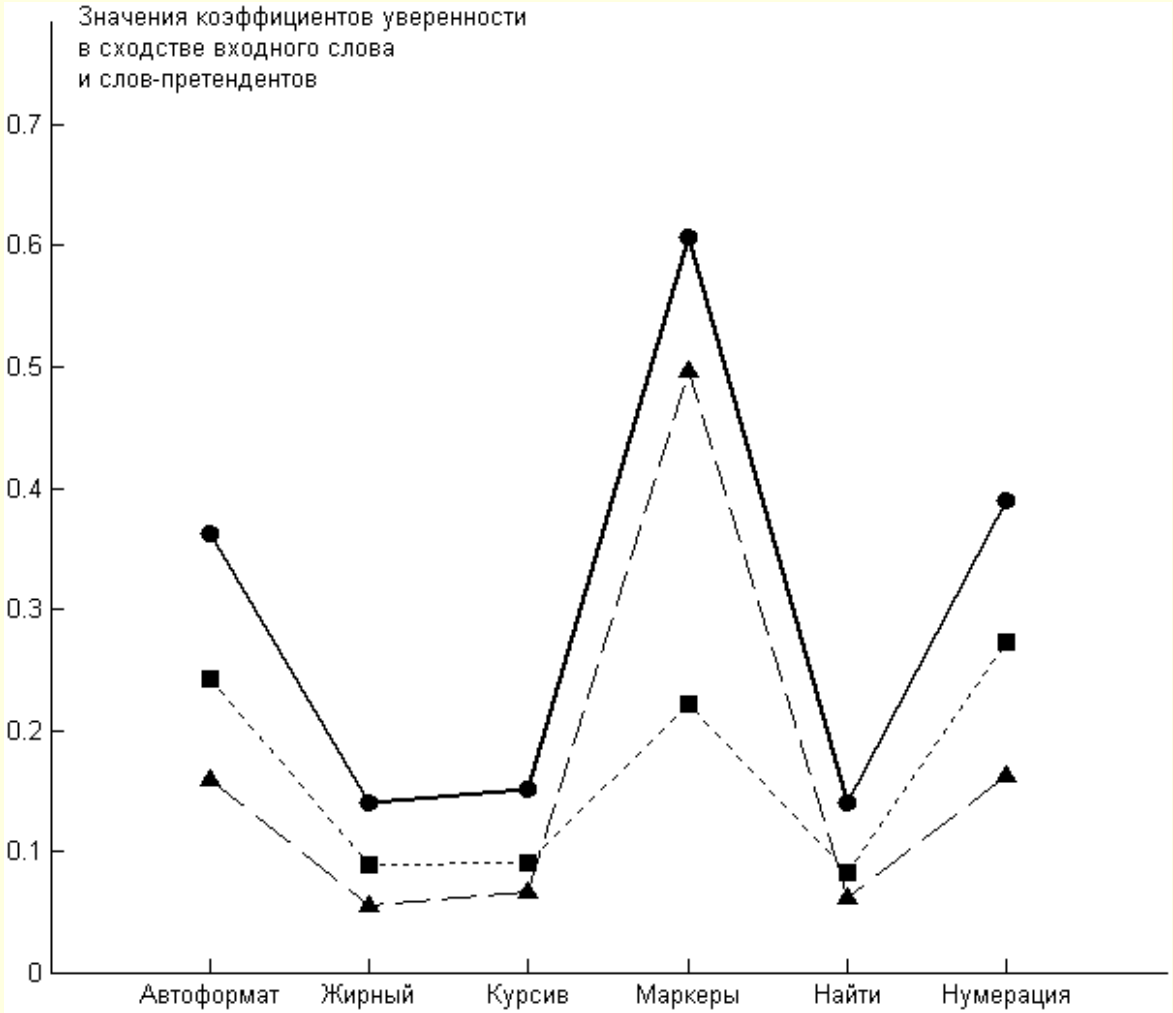
. . . . . .

Схема вывода с учётом ненадёжности знаний

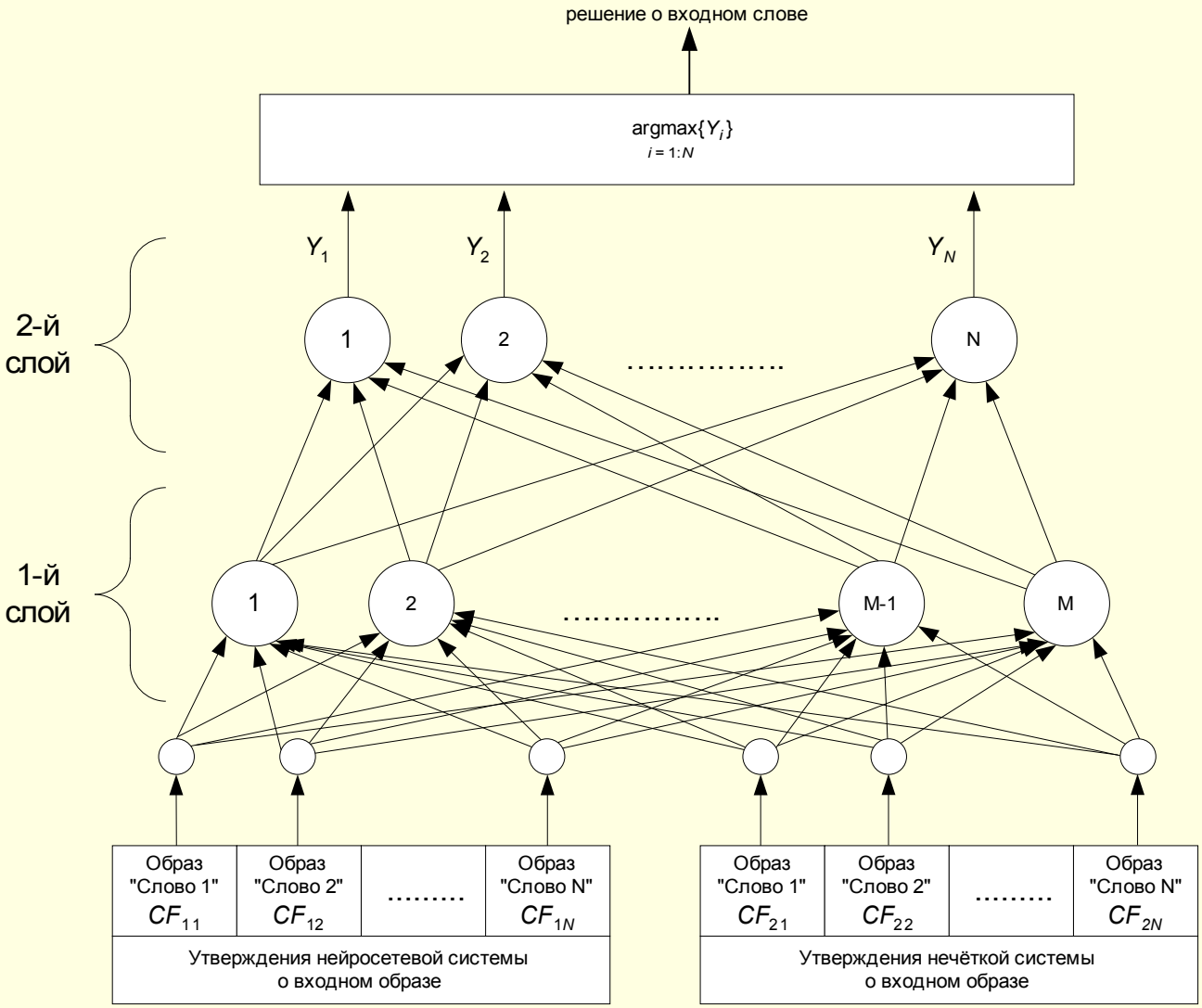


$$CF[V_1(X_1, Y_1)] = CF[V_1, X_1] + CF[V_1, Y_1] - CF[V_1, X_1]^* CF[V_1, Y_1]$$

Результаты коллективного распознавания слова «Маркеры»



Нейросетевой интегратор решений целостного и сегментного каналов



Структура речевой БД для экспериментов по распознаванию

ВОЗРАСТ ДИКТОРА			
ПОЛ ДИКТОРА	18 – 40 лет	41 – 60 лет	∑
мужской	5	1	**6**
женский	2	1	**3**
∑	**7**	**2**	**9**
Каждое из 105 слов произносилось кажжым диктором 3 раза. Всего: 2835 звукозаписей изолированных слов и словосочетаний.			

Методика проведения экспериментальных исследований

При однодикторном распознавании:

- для обучения использовалось по 2 из 3 реализаций каждого слова;
- для тестирования – по 1 реализации.

Эксперименты проводились для каждого из девяти дикторов отдельно, а потом результаты распознавания усреднялись по всем дикторам.

При многодикторном распознавании:

- для обучения использовались все реализации одного женского и четырёх мужских голосов;
- для тестирования – все реализации оставшихся голосов.

Точность многодикторного распознавания

№ диктора	1	2	3	4
Целостный канал	87,6	90,4	63,5	87,6
Сегментный канал	**86,1**	**90,5**	**62,0**	**88,7**
Коэффициенты уверенности	89,5	87,6	72,1	91,4
Нейросеть	89,9	88,1	72,7	91,5

Сравнение двух способов согласования решений каналов

Из таблицы видно, что средняя точность индивидуальной работы целостного и сегментного каналов распознавания составляет **82,27 %** и **81,83 %** соответственно. При объединении каналов в систему коллективного распознавания точность повышается:

− если используется интегратор на основе коэффициентов уверенности, то средняя точность распознавания по всем дикторам повышается до **85,15 %**;

− если нейросетевой интегратор − до **85,55 %**.

Таким образом, алгоритм согласования решений на основе коэффициентов уверенности обеспечивает чуть меньшую точность распознавания, но к его преимуществам можно отнести то, что он не требует обучения.

Декомпозиция многослойной нейронной сети на фрагменты для отображения на многопроцессорную систему

Объектом исследования является нейросетевой аппроксиматор фонем на базе многослойной нейронной сети с полными последовательными связями, состоящей из K слоёв. В такой сети сигналы проходят по слоям последовательно:

$Y^k = F^k(Y^{k-1})$,

где:

$Y^k = \{y_1^k, y_2^k, ..., y_{Nk}^k\}$ — выходной сигнал k-го слоя;

$F(Z) = f^k(A^k \cdot Z + B^k)$ — функциональная модель нейронов k-го слоя;

$k = \overline{1, K}$ — номер слоя;

$A^k = \|a_{ij}^k\|$, $B = \|b_{ij}^k\|$;

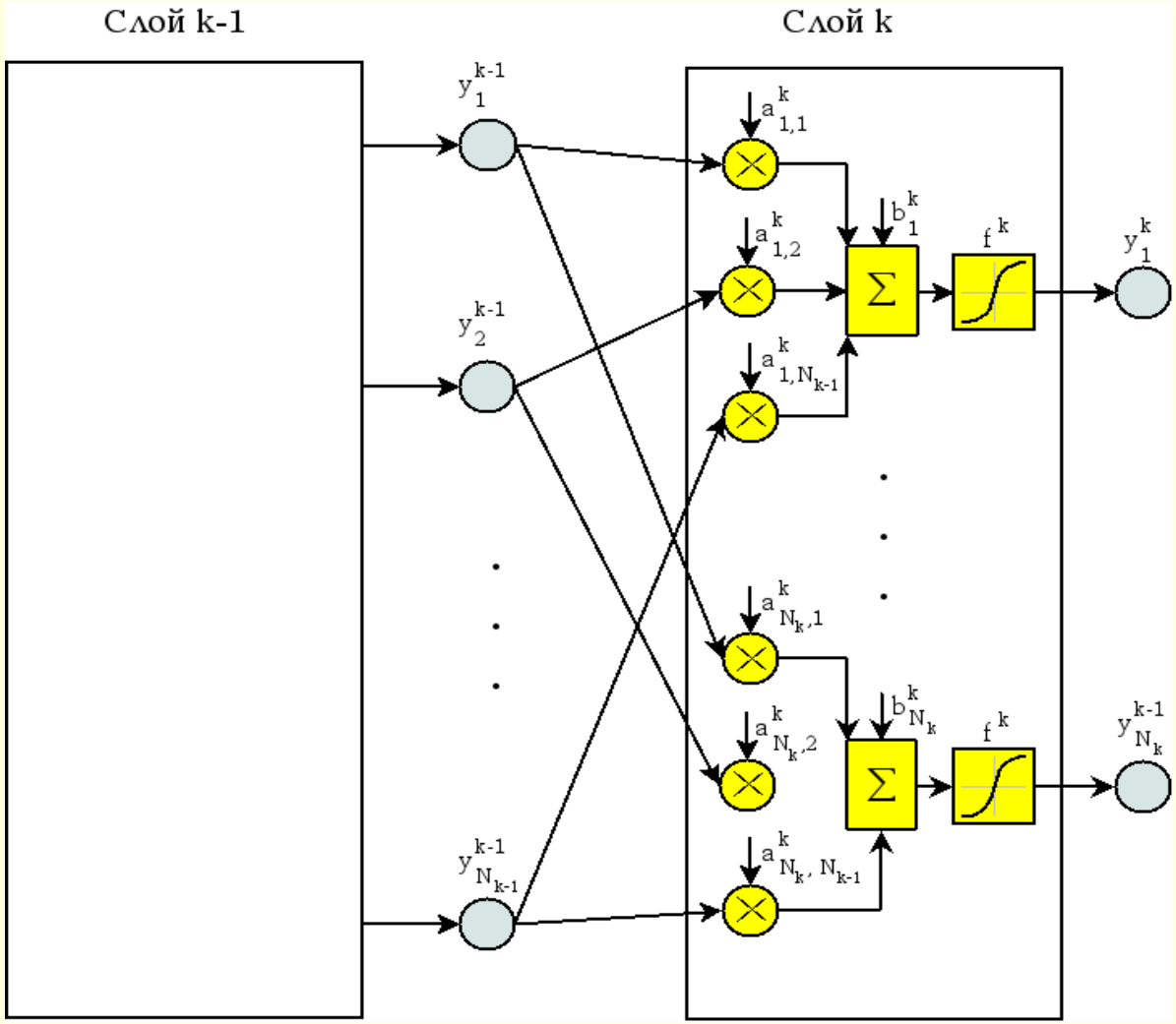
f^k — функция активации нейронов k-го слоя;

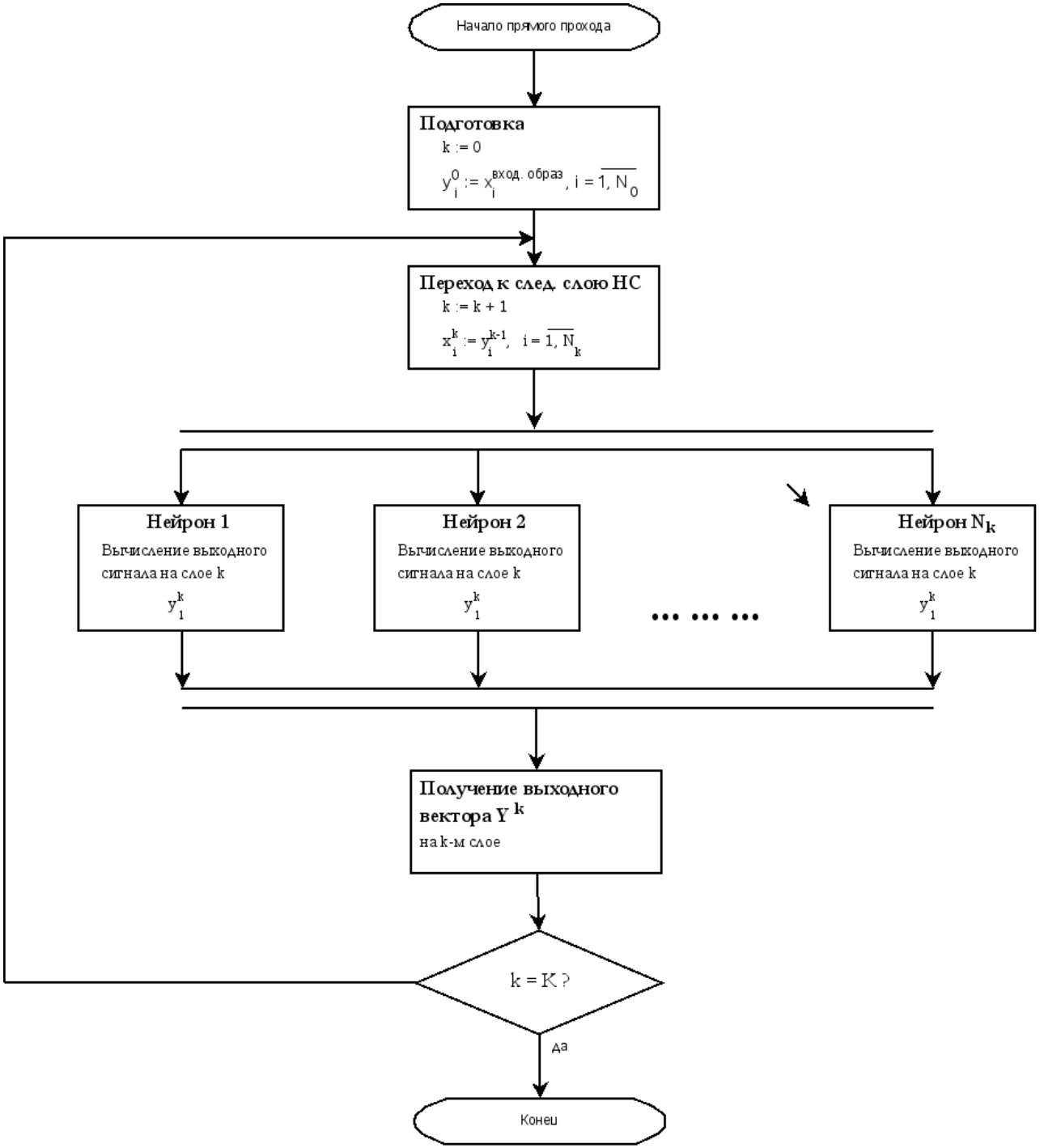
Z — аргумент-вектор, являющийся входным сигналом k-го слоя;

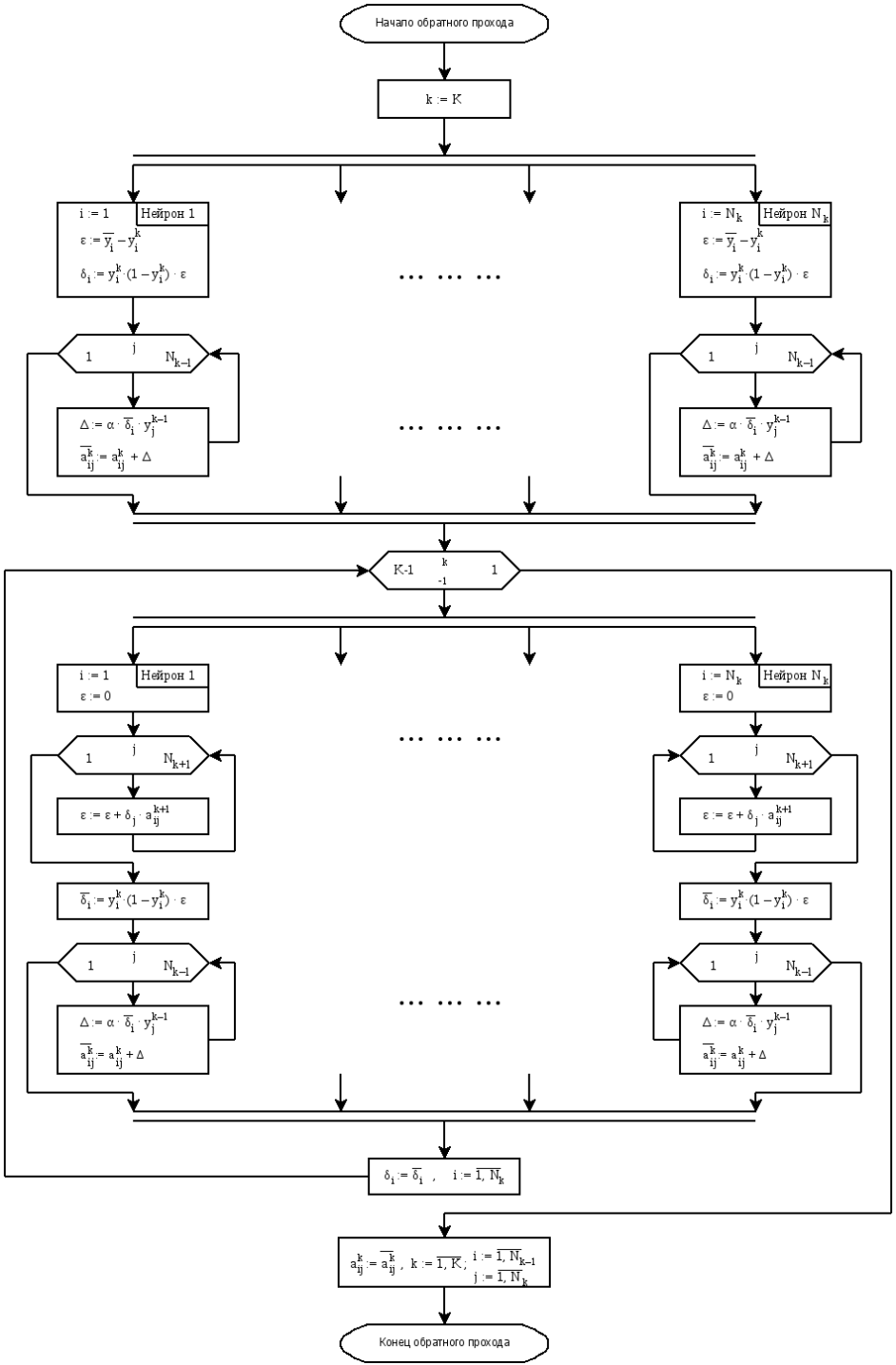
$Y^0 = X$;

$Y = Y^K$ — выходной сигнал нейросети.

Декомпозиция многослойной нейронной сети на фрагменты для отображения на многопроцессорную систему




Начало прямого прохода
Подготовка
k := 0
$y_i^0 := x_i^{\text{вход. образ}}$, $i = \overline{1, N_0}$
Переход к след. слою НС
k := k + 1
$x_i^k := y_i^{k-1}$, $i = \overline{1, N_k}$
Нейрон 1
Вычисление выходного сигнала на слое k
y_1^k
Нейрон 2
Вычисление выходного сигнала на слое k
y_1^k
... ...
Нейрон N_k
Вычисление выходного сигнала на слое k
y_1^k
Получение выходного вектора Y^k
на k-м слое
k = K ?
да
Конец


Начало обратного прохода
k := K
i := 1
Нейрон 1
$\varepsilon := \overline{y}_i - y_i^k$
$\delta_i := y_i^k (1 - y_i^k) \cdot \varepsilon$
j
1
N_{k-1}
$\Delta := \alpha \cdot \overline{\delta}_i \cdot y_j^{k-1}$
$\overline{a}_{ij}^k := a_{ij}^k + \Delta$
$i := N_k$
Нейрон N_k
K-1
k
-1
1
$\varepsilon := 0$
$\varepsilon := \varepsilon + \delta_j \cdot a_{ij}^{k+1}$
N_{k+1}
$\overline{\delta}_i := y_i^k (1 - y_i^k) \cdot \varepsilon$
$\delta_i := \overline{\delta}_i$, $i := \overline{1, N_k}$
$a_{ij}^k := \overline{a}_{ij}^k$, $k := \overline{1, K}$; $i := \overline{1, N_{k-1}}$ $j := \overline{1, N_k}$
Конец обратного прохода

Мультипроцессорная графическая плата как основа кластерной системы

Компьютерный эмулятор нейронной сети построен на кластерной системе с архитектурой, представляющей собой персональный компьютер с:

1) **центральным процессором Intel Core 2 Duo E8500**;

2) **многопроцессорной графической платой nVidia GeForce 9500 GT**.

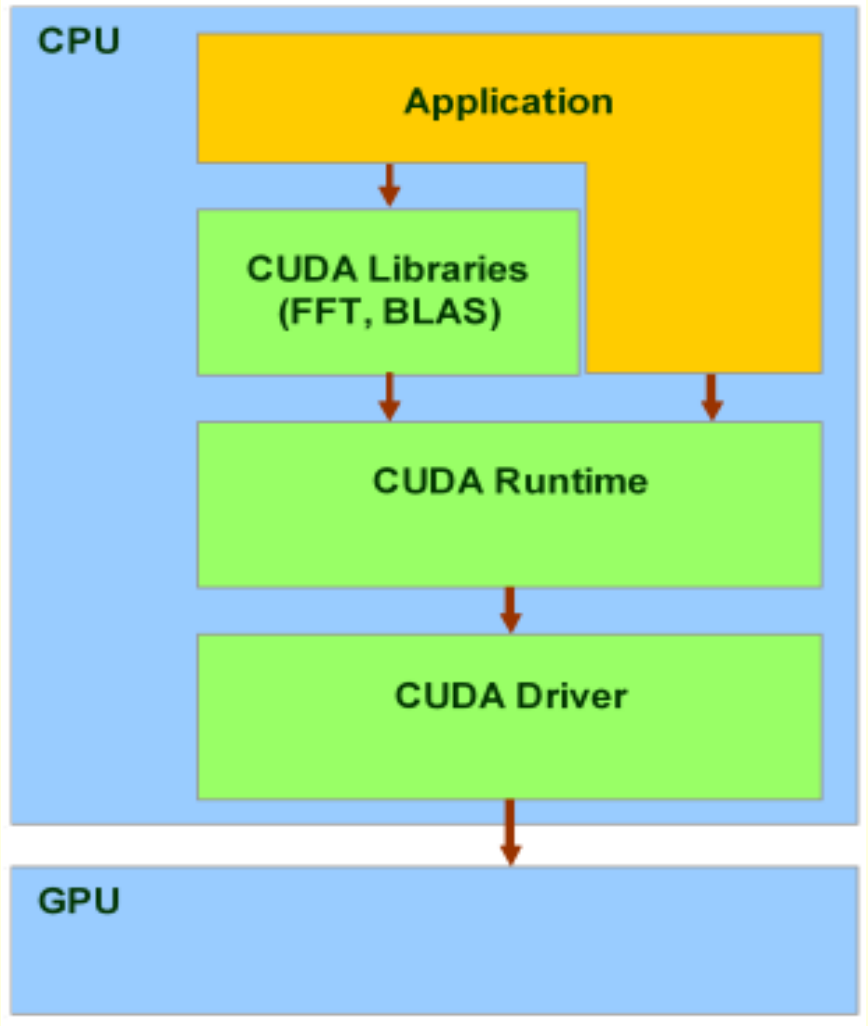
Основные характеристики платы nVidia GeForce 9500 GT

ХАРАКТЕРИСТИКА	КОЛИЧЕСТВЕННОЕ ЗНАЧЕНИЕ
Stream-процессоры	32
Частота ядра, МГц	550
Частота шейдерного блока, МГц	1400
Частота памяти, МГц	800
Объём памяти, Мб	256
Интерфейс памяти	128-bit
Полоса пропускания памяти, Гб/с	25,6
Скорость наложения текстур, млрд./сек	8,8

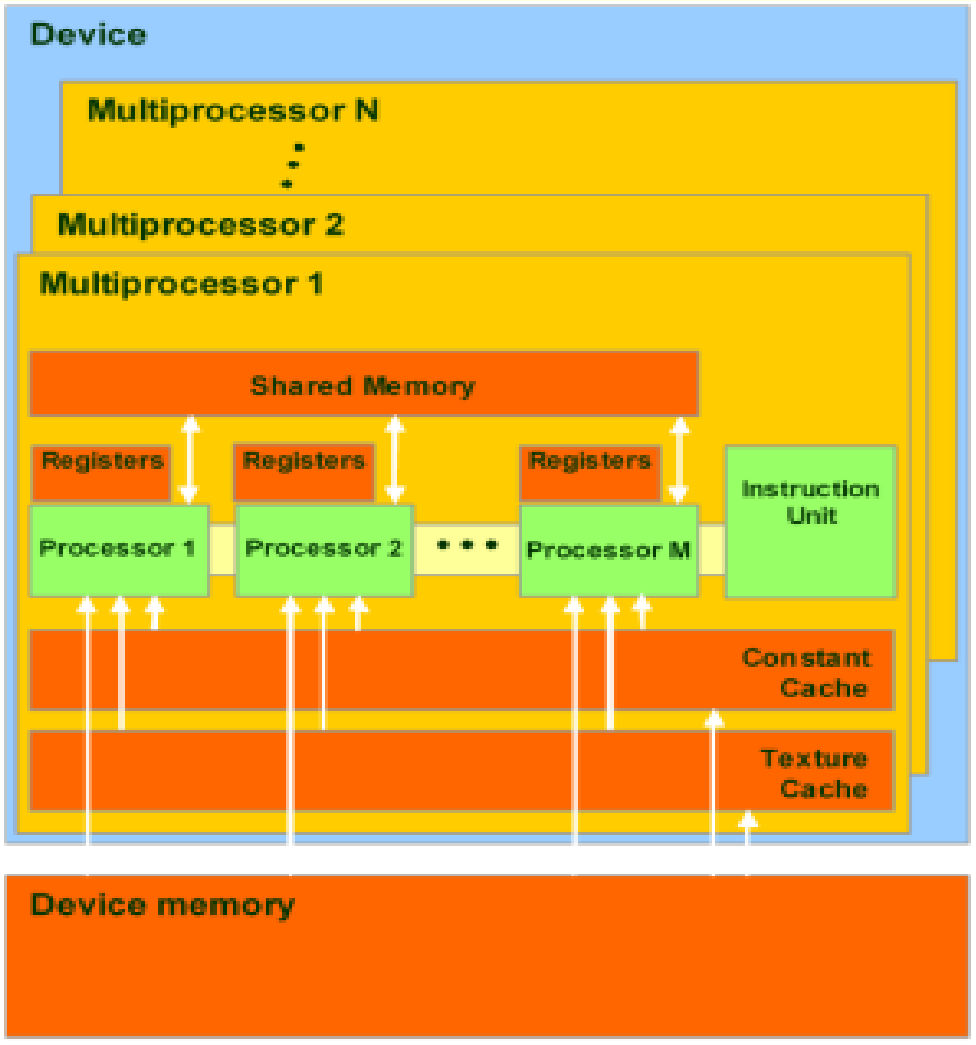
Основные характеристики процессора Intel Core 2 Duo E8500

ХАРАКТЕРИСТИКА	КОЛИЧЕСТВЕННОЕ ЗНАЧЕНИЕ
Тактовая частота, ГГц	3,16
Кэш-память второго уровня, Кб	6144
Кэш-память третьего уровня, Кб	—
Количество ядер	2
Частота системной шины, МГц	1333

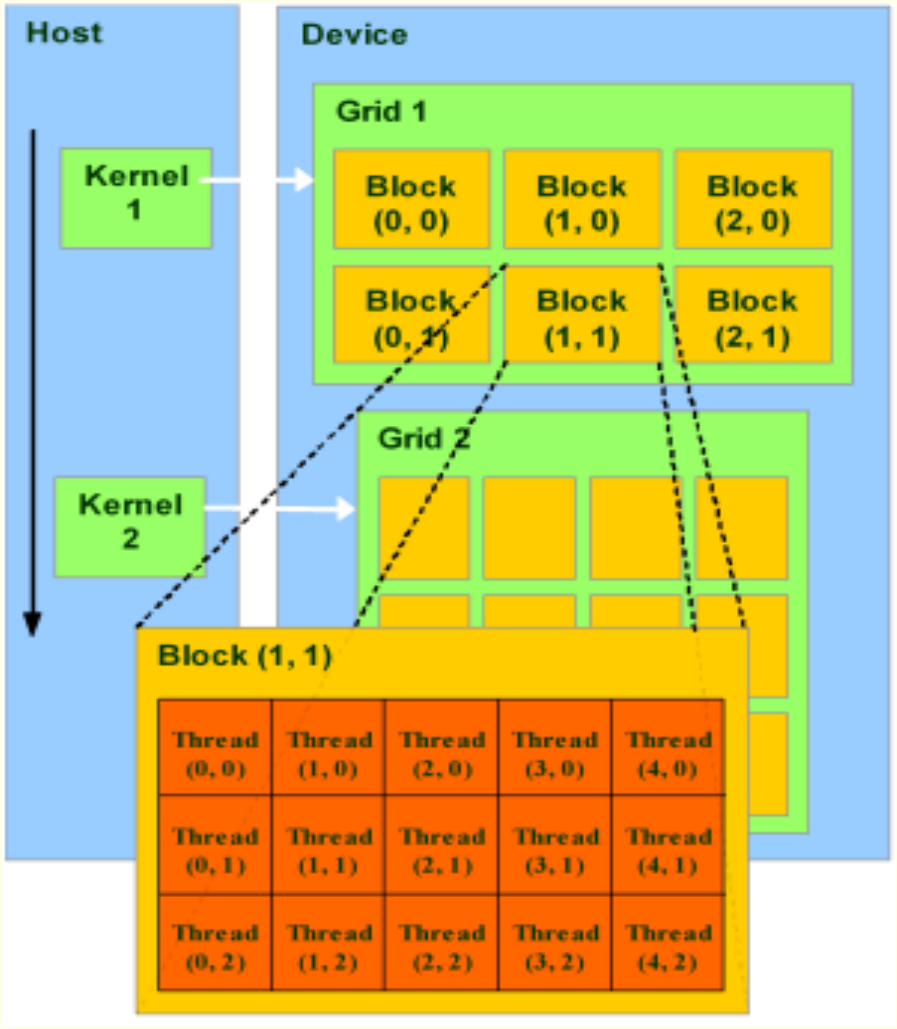
Состав NVIDIA CUDA



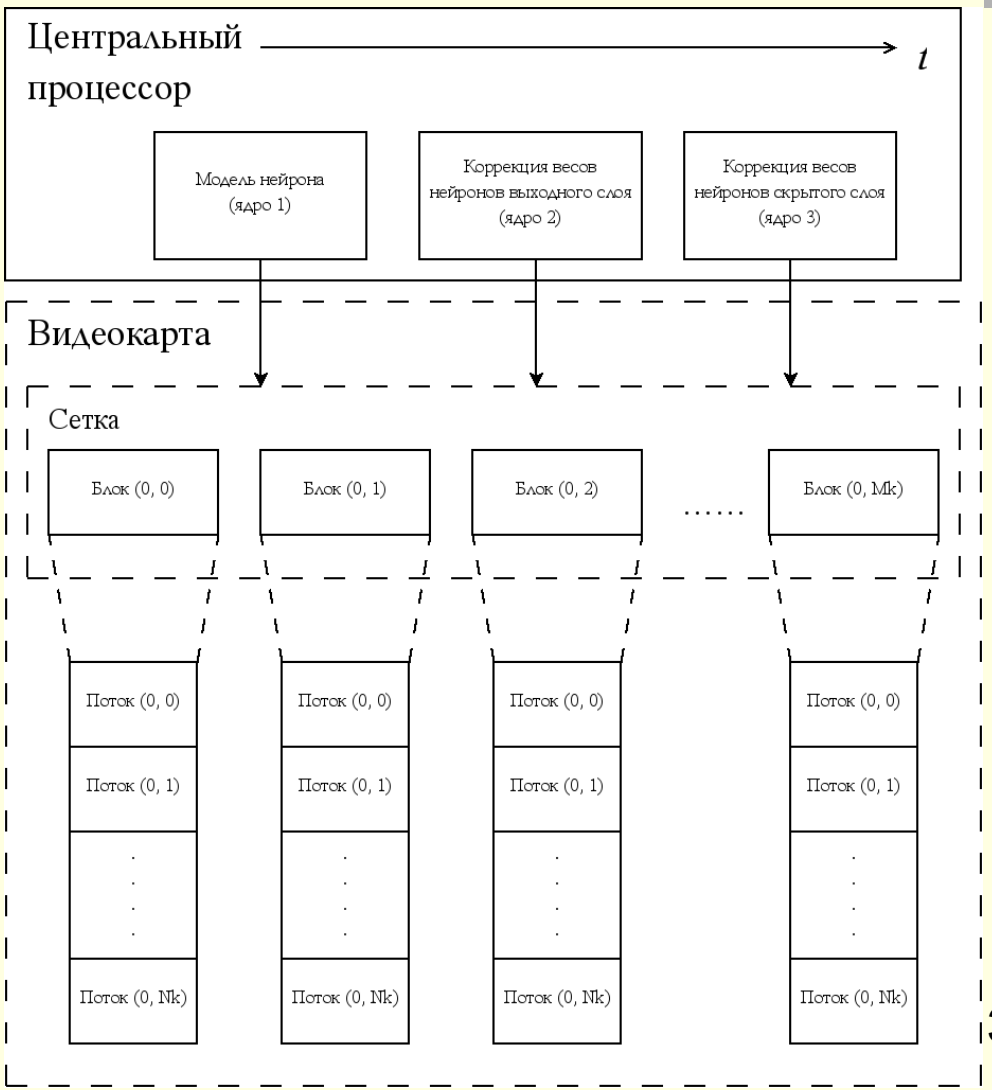
Модель памяти CUDA



Модель программирования CUDA



Отображение потоков на программно-аппаратную вычислительную архитектуру CUDA



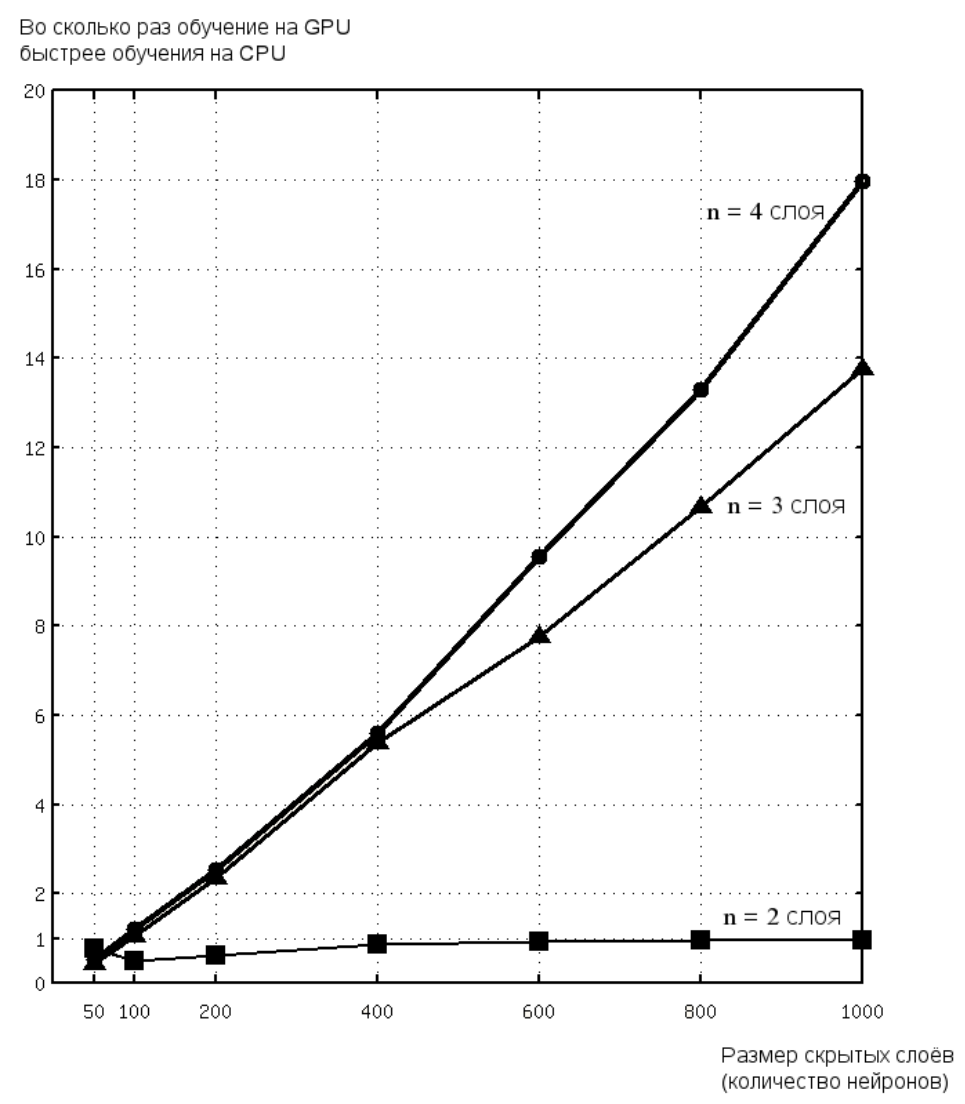
Экспериментальная оценка распараллеливания нейроалгоритмов распознавания и обучения

В экспериментах использовались следующие модели нейросетевого аппроксиматора:

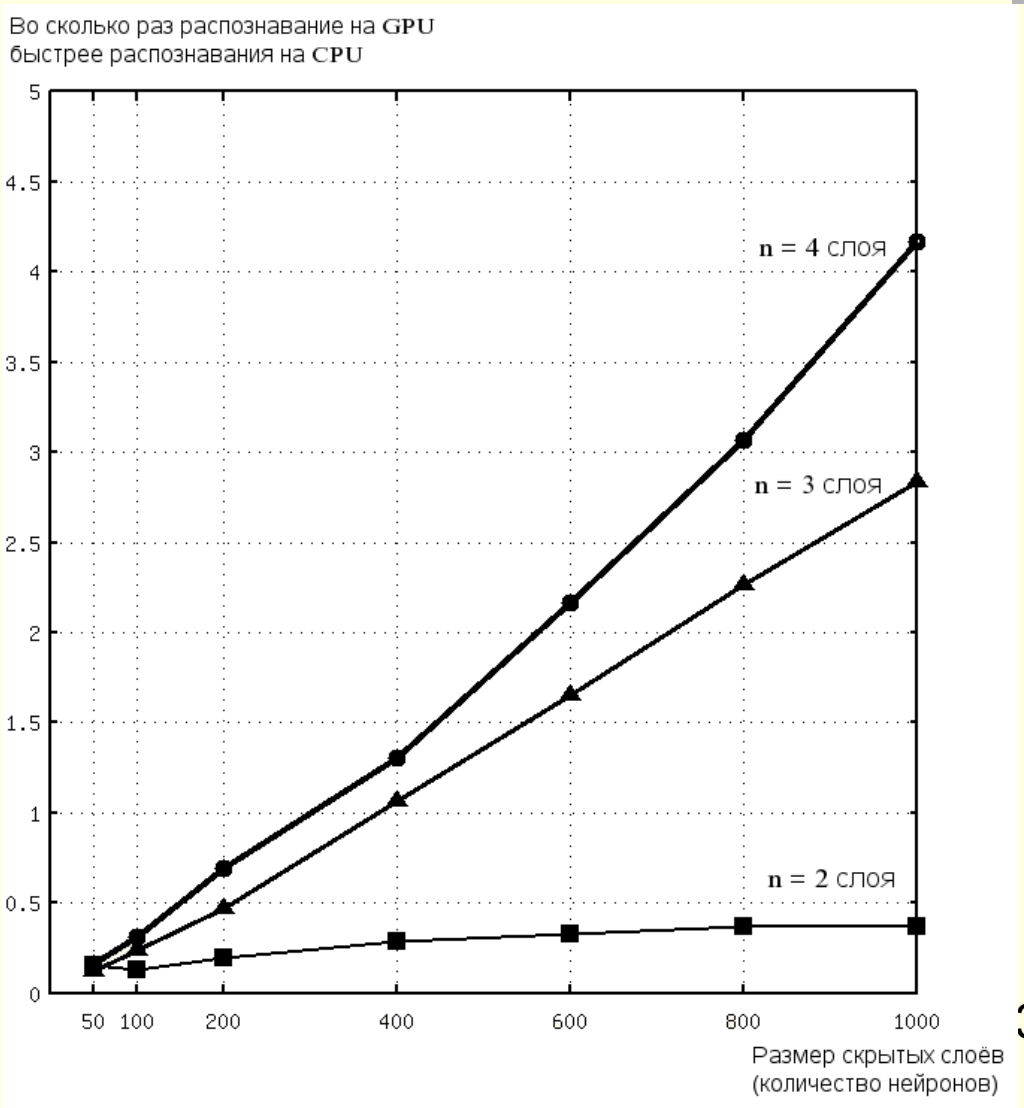
1) параллельная реализация нейроалгоритма на графической плате nVidia GeForce 9500 GT по технологии nVidia CUDA;

2) последовательная реализация нейроалгоритма на центральном процессоре Intel Core 2 Duo E8500.

Ускорение процесса обучения моделей многослойных нейросетей на GPU по отношению к CPU



Ускорение процесса распознавания гласных фонем многослойными нейросетями на GPU по отношению к CPU



Выводы

1. Для моделирования сегментного канала распознавания был разработан алгоритм нейросетевого детектирования фонем, основанный на мел-частотном кепстральном анализе и применении коллектива нейросетевых детекторов фонем. Каждый из детекторов вычисляет темпоральную функцию принадлежности звукового сигнала заданной фонеме. Итоговое решение о распознанном слове принимается в результате согласования вычисленной фонемной цепочки с эталонными транскрипциями с помощью динамического программирования.

Выводы

2. Разработаны алгоритмы согласования решений сегментного и целостного каналов в двухканальной модели, один из которых основан на методе коэффициентов уверенности, а другой – на искусственных нейронных сетях. Объединение решений каналов осуществляется интегратором, представляющим собой нейросетевую экспертную систему или экспертную систему, которая выполняет логический вывод решения на основе ненадёжных знаний. Эксперименты показали, что оба разработанных алгоритма функционирования интегратора позволяют эффективно разрешать конфликтные ситуации, возникающие при расхождении мнений экспертов (каналов распознавания), и повышают точность распознавания в многодикторном режиме на 8 – 9 %. При этом согласование решений на основе коэффициентов уверенности обеспечивает чуть меньшую точность, но не требует обучения.

Выводы

3. Для ускорения процессов нейросетевого распознавания речи и обучения такому распознаванию авторами предложена декомпозиция нейроалгоритма, позволяющая распараллелить его выполнение средствами мультипроцессорной графической платы на основе технологии nVidia CUDA. Были проведены экспериментальные исследования, направленные на оценку эффективности данного распараллеливания по критерию производительности. Эти исследования показали, что параллельная реализация на графической плате позволяет ускорить работу нейросетевого аппроксиматора фонем в несколько раз, как в режиме обучения, так и в режиме распознавания. Наблюдается прямая линейная зависимость выигрыша в производительности, достигаемого путём параллельной реализации нейросетевого аппроксиматора фонем, от сложности его структуры.

Спасибо за внимание