3Б

УДК 004.522:004.934

И.Ю. Бондаренко, О.И. Федяев, К.К. Титаренко
Донецкий национальный технический университет, г. Донецк, Украина
fedyaev@r5.dgtu.donetsk.ua, bond005@yandex.ru, i.const.t@gmail.com

Нейросетевой распознаватель фонем русской речи на мультипроцессорной графической плате

В статье рассматривается применение программно-аппаратных средств современной графической платы с параллельной архитектурой для построения на ней нейросетевого распознавателя фонем русской речи. Предложена декомпозиция нейросетевых алгоритмов на фрагменты для отображения их на многопроцессорную вычислительную систему. Проведены экспериментальные исследования, которые показали преимущество распараллеливания нейровычислений на графической плате для решения задачи распознавания речи в реальном масштабе времени.

Введение

В работе оценивается эффективность реализации нейросетевых алгоритмов параллельной обработки речевого сигнала средствами видеокарты, позволяющей решать задачу распознавания речи в реальном масштабе времени. Эти алгоритмы предложены авторами в работе [1] и основаны на сегментной обработке речевого сигнала с целью определения фонемной структуры распознаваемого речевого слова. Основными преимуществами такого подхода являются позиционная независимость и нечувствительность к изменению временной структуры сигнала.

Локализация и распознавание фонем осуществляется параллельно работающими нейросетевыми аппроксиматорами, которые реализуют модели соответствующих фонем (рис. 1). Аппроксиматоры фонем построены на нейросетях типа «многослойный персептрон» благодаря их универсальным аппроксимирующим свойствам и наличию хороших алгоритмов обучения [2]. Такая структура за счёт настройки каждой нейросети на распознавание одной фонетической единицы даёт возможность существенно снизить вычислительные затраты при обучении нейросетевого ансамбля.

Программная реализация сегментного канала распознавания на основе нейросетевой аппроксимации фонем выявила необходимость большого объёма вычислений при распознавании изолированных слов. В частности, распознавание речевых команд из словаря объёмом 60 слов на одноядерном компьютере с процессором AMD-K6-3DNow (тактовая частота 500 МГц) выполняется в среднем около 6 с, что неприемлемо для организации диалога в реальном времени. Анализ распределения временных затрат при распознавании на компьютере с указанным процессором показал, что 99% вычислений приходится на определение мер близости с помощью искусственных нейронных сетей [3]. В связи с этим представляется актуальной реализация нейросетевых фонемных аппроксиматоров, распараллеленных на логическом уровне, на вычислительной системе с параллельной архитектурой. Сегодня по стоимости наиболее доступной из таких систем является современная графическая плата, поэтому **цель данной работы** – оценить эффективность реализации

нейросетевых алгоритмов распознавания речи на коммерчески доступной графической плате по критерию производительности.

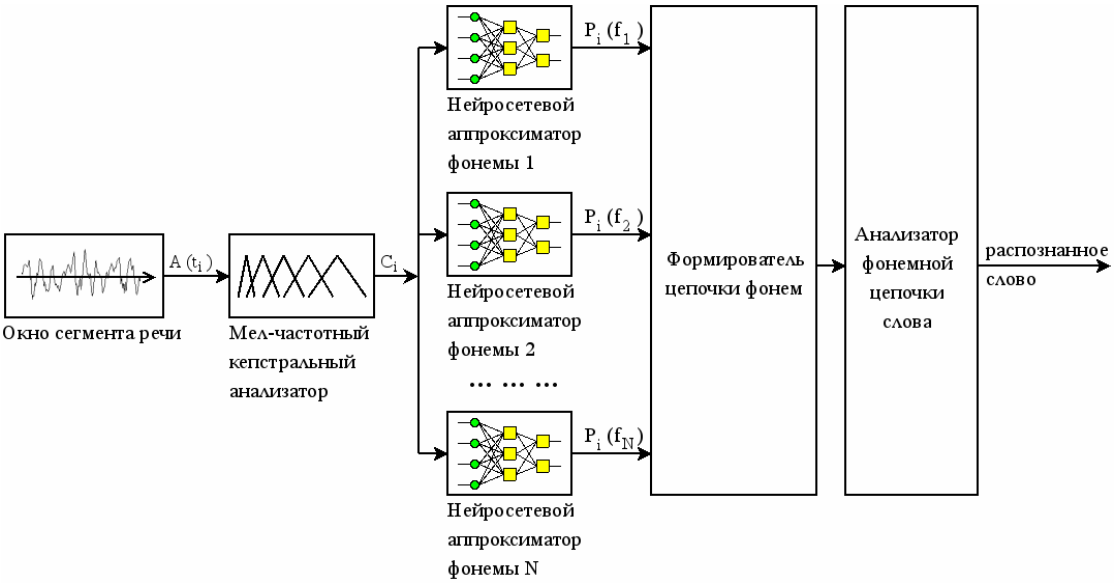

Рисунок 1 – Структура системы распознавания речевых команд на основе нейросетевой аппроксимации фонем

Декомпозиция многослойной нейронной сети на фрагменты для отображения на кластерную 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.

Ставится задача ускорить работу нейросетевого аппроксиматора фонем за счёт распараллеливания процессов обучения и распознавания для различных параметров нейроалгоритма.

Анализ нейровычислений на уровне аппроксиматора показывает, что распараллеливание возможно только в пределах слоя путём независимой реализации функций нейронов, описываемой матричными операциями $A^k \times Z + B^k$ и функцией f^k. Опираясь на это, была выполнена декомпозиция нейросети на фрагменты для отображения её на кластерную систему. Блок-схема распараллеливания прямого хода нейросети (вычисления выходных сигналов всех слоёв) приведена на рис. 2, а блок-схема распараллеливания обратного хода (корректировки синаптических коэффициентов по алгоритму обратного распространения ошибки [4]) – на рис. 3.

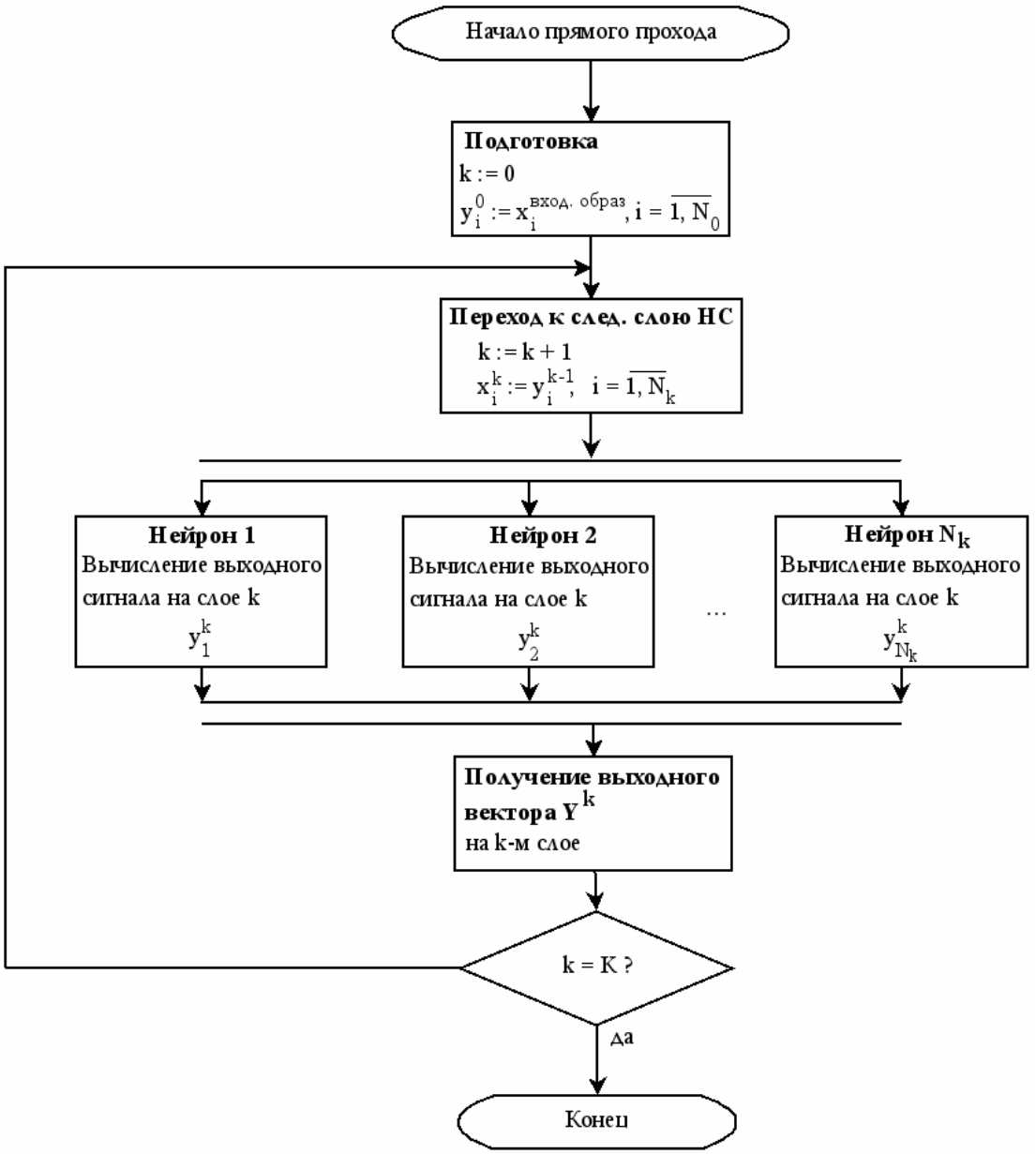

Рисунок 2 – Блок-схема распараллеливания прямого хода в нейросети

Мультипроцессорная графическая плата как основа кластерной системы

Компьютерный эмулятор нейронной сети построен на кластерной системе с архитектурой, представляющей собой персональный компьютер с центральным процессором Intel Core 2 Duo E8500 и многопроцессорной графической платой nVidia GeForce 9500 GT. Организация параллельных вычислений на данной кластерной системе реализована по технологии nVidia CUDA [5]. Декомпозиция нейросетевых вычислений, которая выполнена описанным в предыдущем разделе способом, позволила выделить параллельные вычислительные потоки для отдельных процессорных узлов графической платы. Отображение потоков на программно-аппаратную вычислительную архитектуру CUDA показано на рис. 4.

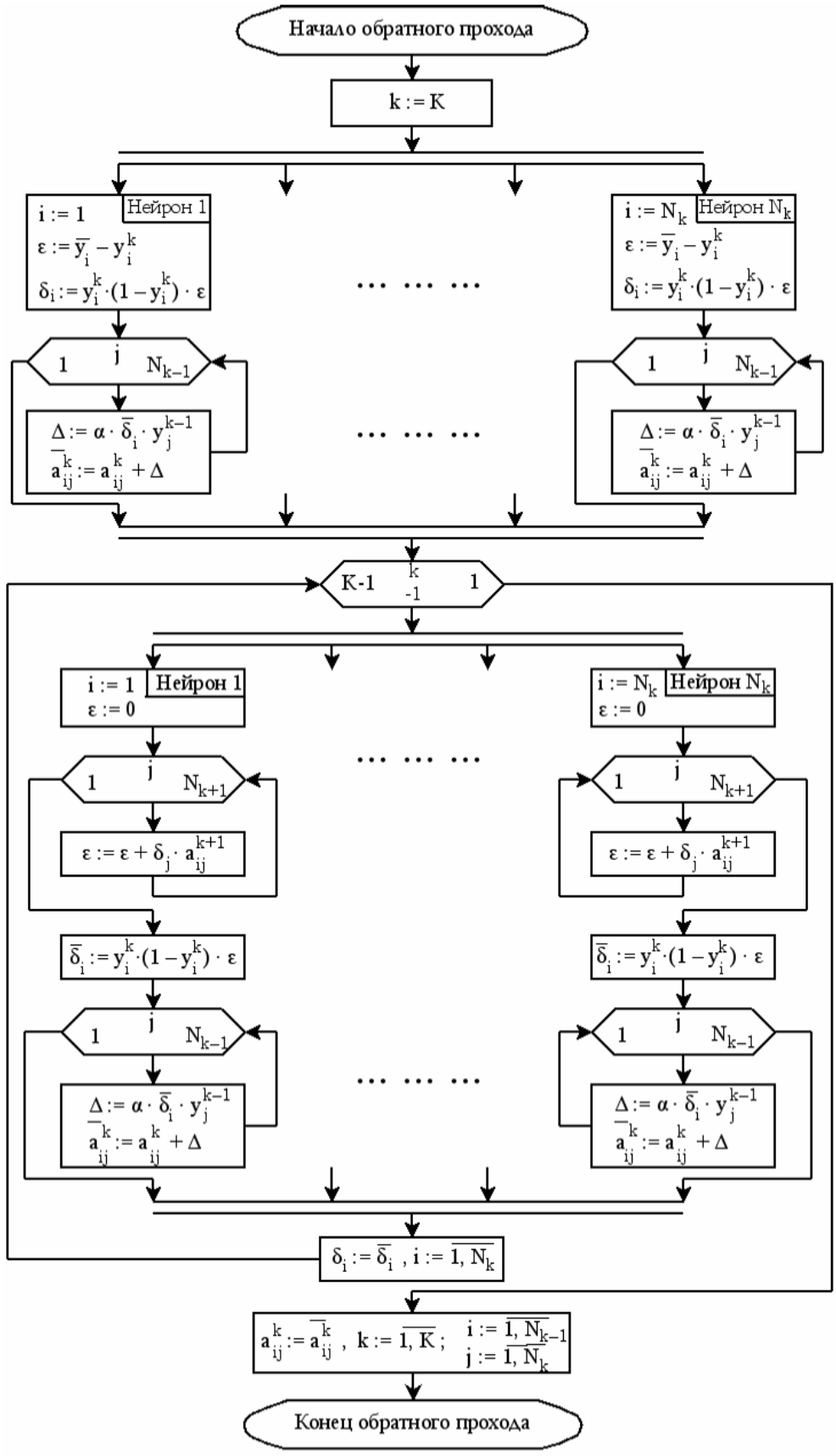

Рисунок 3 – Блок-схема распараллеливания обратного хода в нейросети

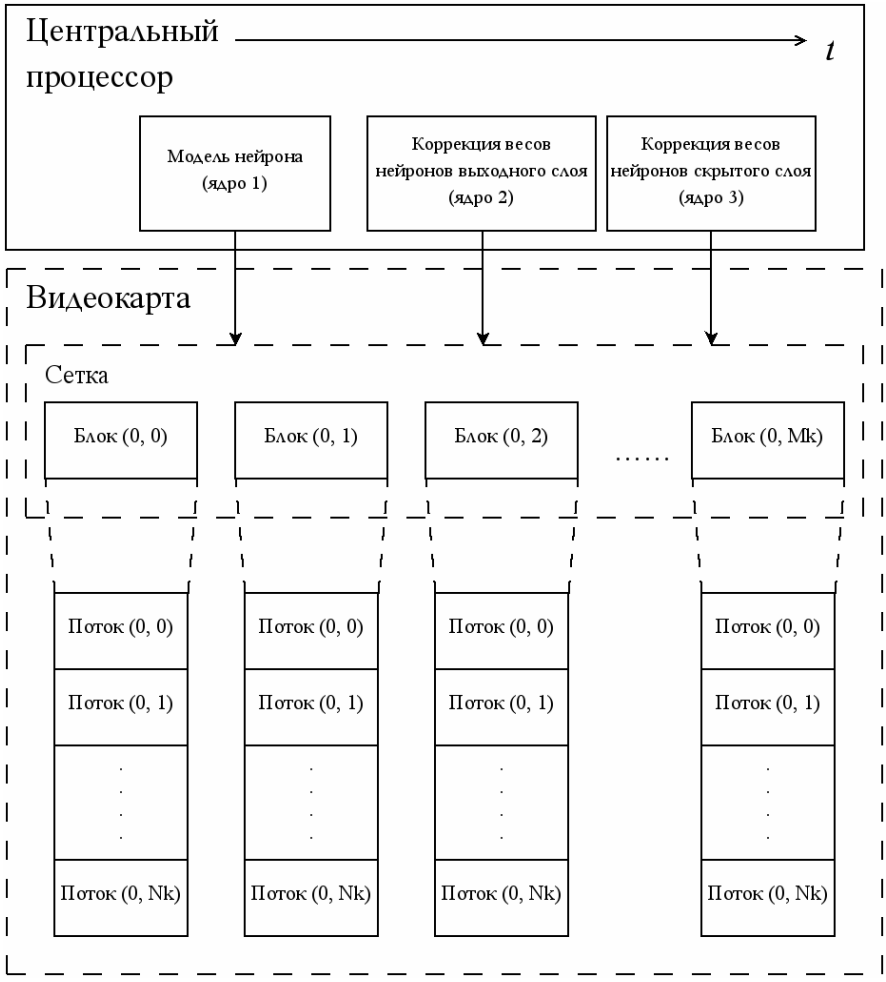

Рисунок 4 – Связь CPU с GPU и детализация группировки потоков в GPU

Характеристики графической платы и центрального процессора, которые применялись при моделировании, приведены в табл. 1 и табл. 2 соответственно.

Таблица 1 – 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/с	8,8

Таблица 2 – 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

Экспериментальная оценка распараллеливания нейроалгоритмов распознавания и обучения

Для того чтобы оценить производительность нейросетевых аппроксиматоров, распараллеленных на графической плате (GPU), по сравнению с реализацией на центральном процессоре (CPU), были проведены испытания данных нейросетевых аппроксиматоров на двух задачах: 1) обучения распознаванию гласных фонем русского языка и 2) собственно распознавания этих фонем.

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.

Обучающие и тестовые данные формировались на основе созданной авторами речевой базы данных, включавшей в себя 6 изолированно произнесённых фонем «А», «О», «У», «Э», «И», «Ы», записанных восьмибитными отсчётами в формате WAV PCM с частотой 11025 Гц.

Формирование обучающего и тестового множеств осуществлялось по методу Windowing с длиной скользящего окна 15 мс и перекрытием при скольжении 5 мс. Вырезаемый окном фрагмент речевого сигнала подвергался мел-частотному кепстральному анализу [6], в результате чего формировалось 13 мел-частотных кепстральных коэффициентов. Чтобы учесть речевой контекст и эффект коартикуляции, входной сигнал нейросетевого аппроксиматора включал в себя не только информацию из текущего окна, но также информацию из двух предыдущих и двух последующих окон. Таким образом, размер входного сигнала нейросети составлял 65 элементов, а размер выходного сигнала – 6 элементов (по числу распознаваемых фонем).

Сравнительные результаты работы последовательной (на CPU) и параллельной (на GPU) реализаций нейросетевых аппроксиматоров фонем различной структуры приведены в табл. 3, 4.

Таблица 3 – Эффективность реализации процесса обучения нейросетевых аппроксиматоров на GPU по отношению к CPU

Количество нейронов в скрытом слое	Выигрыш в скорости для аппроксиматоров с разной структурой, раз		
	1 скрытый слой	2 скрытых слоя	3 скрытых слоя
50	0,76	0,44	0,54
100	0,50	1,05	1,21
200	0,63	2,37	2,53
400	0,85	5,40	5,60
600	0,92	7,75	9,56
800	0,95	10,65	13,30
1000	0,97	13,76	17,97

Ускорение, достигаемое распараллеливанием на графической плате процесса обучения нейросетевого аппроксиматора фонем, показано на рис. 5. Можно видеть, что чем более сложную структуру имеет нейросетевой фонемный аппроксиматор, тем более эффективным становится его распараллеливание, причём эта зависимость является линейной.

Аналогичную картину можно наблюдать и при распараллеливании процесса распознавания фонем (рис. 6).

Таблица 4 – Эффективность реализации процесса распознавания фонем на GPU по отношению к CPU

Количество нейронов в скрытом слое	Выигрыш в скорости для нейросетевых фонемных аппроксиматоров с разной структурой, раз		
	1 скрытый слой	2 скрытых слоя	3 скрытых слоя
50	0,15	0,12	0,16
100	0,13	0,24	0,31
200	0,20	0,47	0,69
400	0,29	1,06	1,31
600	0,33	1,65	2,17
800	0,37	2,27	3,07
1000	0,37	2,83	4,17

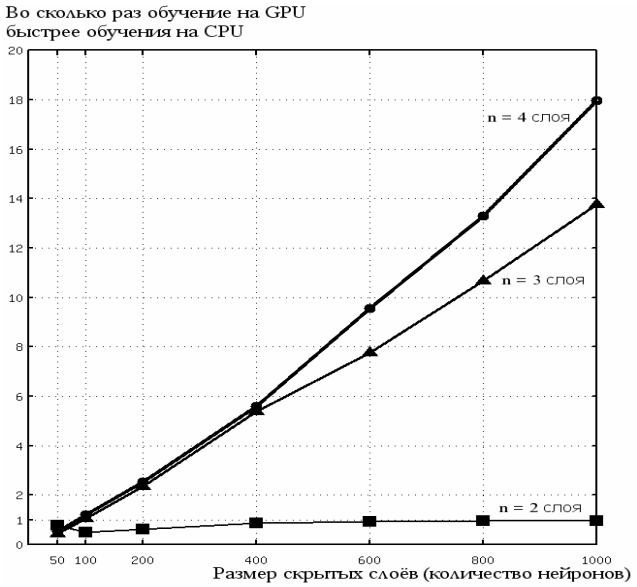

Рисунок 5 – Ускорение процесса обучения моделей многослойных нейросетей на GPU по отношению к CPU

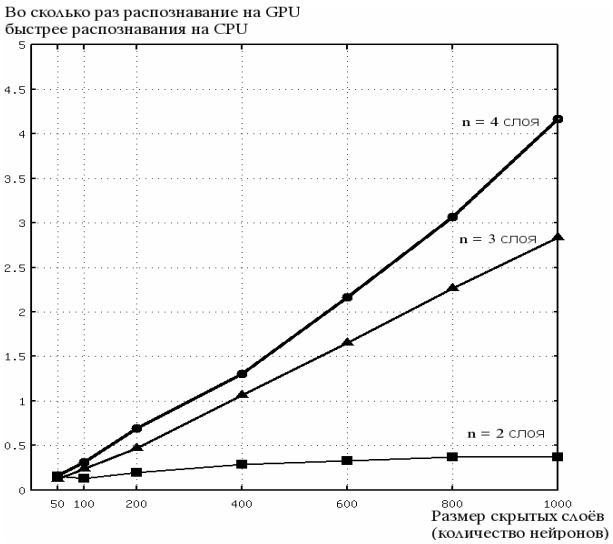

Рисунок 6 – Ускорение процесса распознавания гласных фонем многослойными нейросетями на GPU по отношению к CPU

Выводы

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.

Литература

1. Бондаренко И.Ю. Сегментно-целостная структура канала речевого управления программными системами / И.Ю. Бондаренко, С.А. Гладунов, О.И. Федяев // Сб. трудов нац. конф. по искусств. интеллекту с междунар. участием КИИ-2006. – М. : Физматлит, 2006. – С. 841-849.
2. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / А.Н. Горбань // Сибирский журнал вычислительной математики. – 1998. – Т. 1, № 1. – С. 12-24.
3. Гладунов С.А. Аппаратно-программные средства раздельной локализации фонем в системах речевого взаимодействия человека с ЭВМ : автореф. дисс. канд. техн. наук: 05.13.13 / С.А. Гладунов. – ДонНТУ. – Донецк, 2005. – 22 с.
4. Y. LeCun. Efficient BackProp / Y. LeCun, L. Bottou, G. Orr and K. Muller // in Orr, G. and Muller K. Neural Networks: Tricks of the trade. – New-York : Springer, 1998. – P. 5-50.
5. Jason Sanders. CUDA by Example: An Introduction to General-Purpose GPU Programming / Jason Sanders, and Edward Kandrot. – Reading, Massachusetts: Addison-Wesley Professional, 2010. – 312 p.
6. Nelson Morgan. Automatic Speech Recognition: An Auditory Perspective / Nelson Morgan, Hervé Bourlard and Hynek Hermansky // in Steven Greenberg and William A. Ainsworth. Speech Processing in the Auditory System. – New-York : Springer, 2004. – P. 309-338.

І.Ю. Бондаренко, О.І. Федяєв, К.К. Титаренко

Нейромережний розпізнавач фонем російської мови на багатопроцесорній графічній платі

У статті розглядається застосування програмно-апаратних засобів сучасної графічної плати з паралельною архітектурою для побудови на ній нейромережного розпізнавача фонем російської мови. Запропонована декомпозиція нейромережних алгоритмів на фрагменти для відображення їх на багатопроцесорну обчислювальну систему. Проведені експериментальні дослідження, які показали перевагу розпаралелювання нейрообчислень на графічній платі для вирішення задачі розпізнавання усного мовлення у реальному масштабі часу.

I.Yu. Bondarenko, O.I. Fedyaev, K.K. Titarenko

Classifiers Construction Based on Separate Hyper Surfaces

In the paper the usage of the modern graphics card sw/hw tools with the parallel architecture for construction of the Russian speech neural network phoneme recognizer is considered. Decomposition of neural network algorithms into fragments for its mapping on the multiprocessor system is offered. The carried out experiments showed advantage of the parallelization neurocomputing on a graphics card for automatic speech recognition in real time.

Статья поступила в редакцию 09.07.2010.