# НЕЙРОСЕТЕВОЙ МЕТОД ФОНЕТИЧЕСКОЙ СЕГМЕНТАЦИИ РЕЧЕВОГО СИГНАЛА

**Гладунов С.А., Федяев О.И.**
Кафедра ПМИ, ДонНТУ
gladunov@ukr.net, fedyaev@r5.dgtu.donetsk.ua

### *Abstract*

***Gladunov S.A., Fedyaev O.I. A neural networking method of speech signal phonetic segmentation.*** *A new method proposed of automated speech recognition based on neural approximation of signal. The method is directed to solve a problem of temporal uncertainty in human speech and it deals with phonetic structure of expressions.*

### *Введение*

Задача автоматического распознавания речи до сих пор не имеет качественного решения, которое бы позволило пользователю полноценно взаимодействовать с ЭВМ посредством голоса. Это обусловлено сложной частотно-временной структурой речевого сигнала и его значительной нестабильностью при изменении условий произнесения. Для решения указанной проблемы к настоящему моменту предложено две группы методов:

- параметрические, направленные на математическое преобразование речевого сигнала с выделением и стабилизацией основных информативных признаков (преобразование Фурье, цифровая фильтрация и др. [1]);
- лингвистические, целью которых является контекстная обработка высказывания (методы динамического программирования [2], скрытых Марковских моделей [3] и др.). В качестве исходной информации при этом используются результаты параметризации.

В данной работе описывается нейросетевой алгоритм параметризации речевого сигнала, нечувствительный к длительности произнесения фонем. Результаты работы алгоритма являются исходными данными для лингвистической обработки. Конечной целью работ в этом направлении является реализация модуля ввода речевых команд управления информационными системами [4].

### *1. Постановка задачи*

Задача автоматического распознавания речи может быть поставлена следующим образом. Имеется речевое высказывание $w_i$:

$$w_i = (A(w_i), s(w_i)), \quad w_i \in W \qquad (1)$$

где

$A_j(w_i)$ – j-е произнесение высказывания $w_i$;

$s(w_i)$ – символьное представление информации, содержащейся в высказывании $w_i$, $s(w_i) \in S$; S – словарь;

W – множество допустимых высказываний.

Требуется найти отображение R:

$$\forall w_i \in W, \quad R\,[A_j\,(w_i)] = s\,(w_i). \qquad (2)$$

### *2. Метод нейросетевой аппроксимации фонем*

Для решения этой задачи в работе предложен метод скользящего фонетического анализа, основанный на предположении, что вокализованный речевой сигнал состоит из стационарных участков, характеризующих фонемы, и нестабильных отрезков, относящихся к межфонемным переходам. Причем длительность произнесения слов в основном определяется длительностью стационарных участков, на которых характеристики речевого сигнала достаточно стабильны (рис. 1). Предполагается, также, что стационарные участки однозначным образом характеризуют соответствующие фонемы.

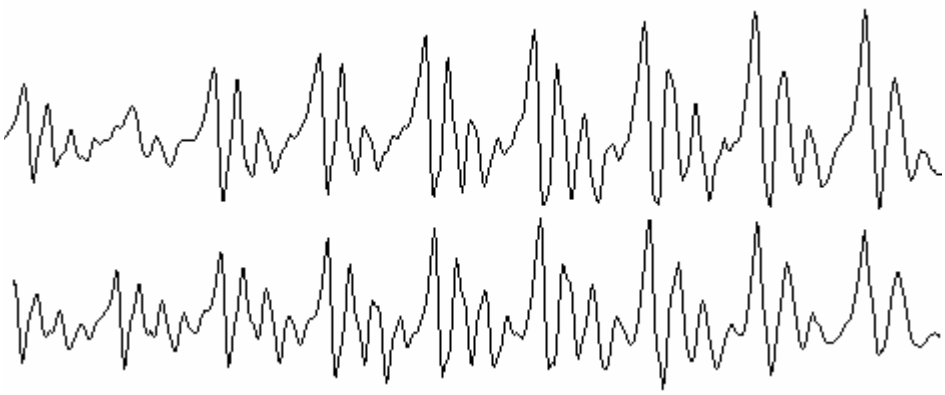

Рис. 1 – Фрагменты двух произнесений фонемы “А”

Сутью метода является определение меры сходства текущего фрагмента речевого сигнала с каждой из фонем для определения наиболее достоверной фонетической цепочки. Опишем метод формально.

Пусть $A_w(t)$ – акустическое представление высказывания w; $F_k(t)$ – акустическое представление некоторой фонемы. Требуется определить, является ли фонема, описываемая $F_k(t)$, фрагментом высказывания $A_W(t)$.

Представим $F_k(t)$ на отрезке $[t_0; t_1]$ в виде множества пар

$$\{(X'(t), Y'(t))\} \qquad (3),$$

где $X'(t) = (F_k(t–m), F_k(t–m+1), \ldots F_k(t–1))$, $m = const$; $Y'(t) = F_k(t)$; $t_0 \le t \le t_1$ Аналогично представим $A_w(t)$ в виде множества пар

$$\{X(t); Y(t)\} \qquad (4).$$

Представление $F_k(t)$ в виде (3) позволяет сформировать нейросетевую функцию

$$NET: NET(X'(t)) = Y'(t). \qquad (5)$$

Тогда меру отличия $Err_k$ участка $A_w(t)$ при $t \in [t_н; t_к]$ от $F_k(t)$ определяется:

$$Err_k(t) = \left/ Y(t) - NET(X(t)) \right/. \qquad (6)$$

Таким образом, получаем новое параметрическое описание исходного сигнала:

$$A_w(t) \rightarrow (Err_1(t), Err_2(t) \ldots Err_n(t)), \qquad (7)$$

где $Err_k(t)$ – мера отличия участка сигнала $A_w(t)$ от k-й фонемы на фрагменте сигнала длительности m.

### *3. Организация системы распознавания речи на основе метода нейросетевой аппроксимации фонем*

При реализации предложенного метода возникает вопрос о стабильности полученных параметров, определяющей конечный результат распознавания. Возможны три ситуации:

1. извлеченная информация непригодна для дальнейшего распознавания;
2. информация достаточна, но требует лингвистического анализа;
3. информация достаточна и не требует дальнейшей обработки.

Для выяснения этого вопроса был проделан ряд экспериментов, направленных на исследование параметрических описаний $A_w(t)$ (7). Результат одного из таких экспериментов приведен на рис. 2.

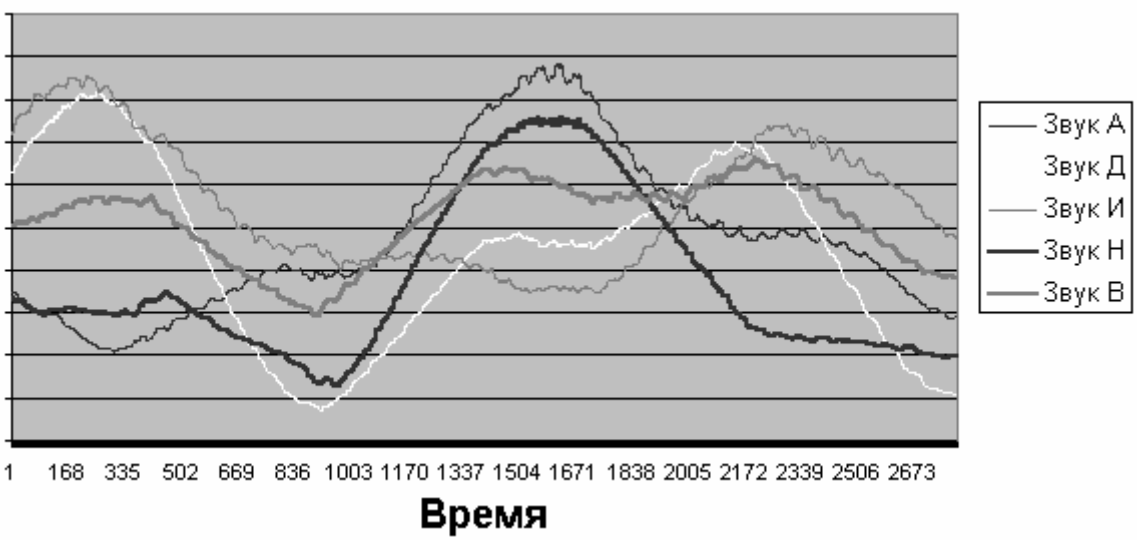


Рис. 2 – Графики мер отличия $Err_k(t)$ фонем “а”, “д”, “и”, “н”, “в” от звуков слова «один»

Представленный график показывает, что полученные параметры информативны, минимумы отличия достигаются на соответствующих фонемах, но при этом имеют место погрешности (такие как звук «н» между «а» и «д» и звук «д» после «н»). В других случаях (было исследовано более 300 вариантов произнесения 60 высказываний) наблюдаются аналогичные результаты. На основании полученных результатов сделан вывод о целесообразности использования метода совместно с методами лингвистического анализа. Сочетание этих двух подходов представлено функциональной схемой на рис.3.

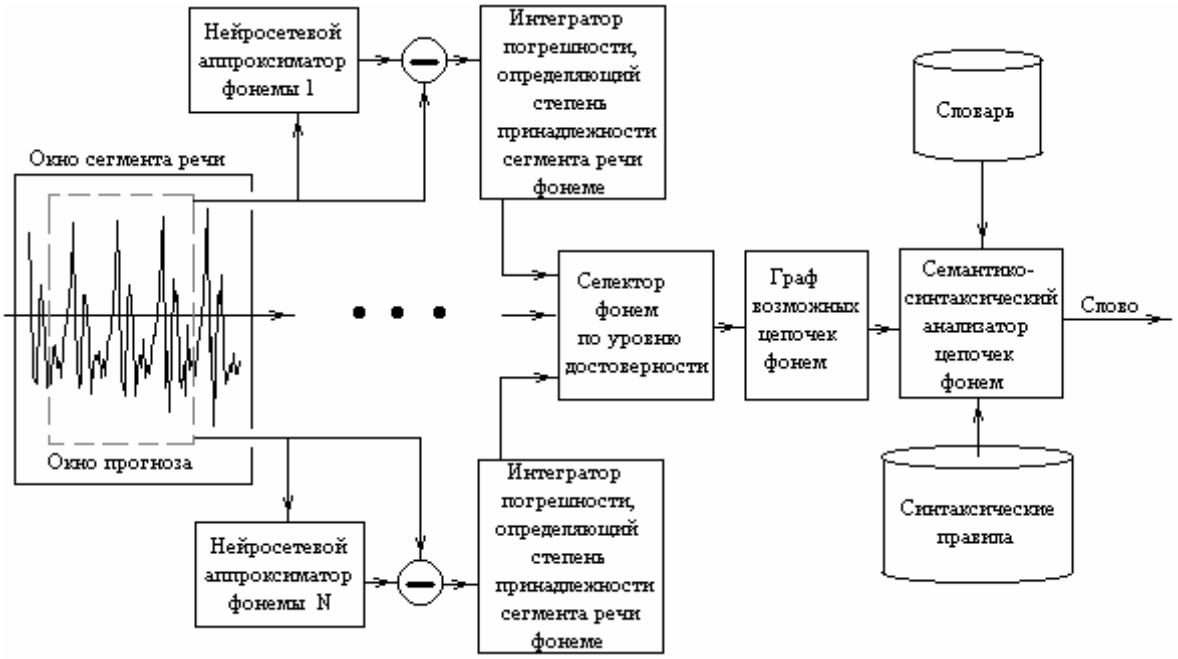


Рис. 3 – Схема распознавания речи на основе нейросетевой аппроксимации фонем

Первый уровень схемы состоит из N нейросетей, каждая из которых обучена на распознавание одной фонемы. При этом несколько различных сетей могут реализовать одну фонему (или её фрагменты). Для всех нейросетей используется один и тот же входной образ, получаемый из «окна прогноза», скользящего вдоль оси времени в пределах анализируемого сегмента речи. Прогнозные значения на выходах нейросетей сравниваются с реальным значением речевого сигнала, и определяется ошибка, характеризующая степень принадлежности текущего окна данной фонеме. На втором уровне схемы полученная ошибка накапливается на всей протяженности окна сегмента речи. Интегральная ошибка для данного сегмента поступает на третий уровень, где из всех фонем селектором выбираются наилучшие по критерию минимума ошибки. Полученный набор фонем участвует в формировании цепочек, представляющих гипотезы о произносимом слове. На этом этапе обработки формируется описание сигнала в виде матрицы достоверности фонем на каждом временном интервале, которая представляет собой исходную информацию для метода динамического прогнозирования, реализующего контекстную обработку. Этот метод позволяет выбрать в словаре цепочку фонем, наиболее близкую к произнесенному высказыванию.

В процессе моделирования были установлено, что в качестве нейросетевых аппроксиматоров лучший результат дают трехслойные нейросети обратного распространения ошибки с 60 входами и распределением нейронов по слоям 20-10-1. Формирование обучающего множества из фонем, выделенных из слов словаря, позволяет добиться большей точности распознавания по сравнению с использованием отдельно произносимых фонем. При этом эксперименты показали, что стабильного распознавания удается достичь только при работе с вокализованными фонемами [5], поэтому был разработан и применен дополнительный алгоритм сегментации речевого сигнала на вокализованные и невокализованные участки по уровню его энергии:

Шаг 1. Вычисляется энергия сигнала S(t) как набор сумм абсолютных значений амплитуды сигнала в рамках скользящего окна заданного размера N;

Шаг 2. Определяются множества точек $B=\{b_1, b_2, \dots b_n\}$ и $K=\{ k_1, k_2, \dots k_n \}$:

$$\begin{cases} \text{если } S(t)>P \text{ и } S(t_1)<P, \text{ то } b_i = t; \\ \text{если } S(t)<P \text{ и } S(t_1)>P, \text{ то } k_i = t_1,\ i=i+1 \end{cases}$$

для всех значений t, где P=const – некоторый заранее определенный порог;

Шаг 3. Среди точек из B и K в качестве границ вокализованных интервалов выбираются только те точки $b_i$ и $e_i$, для которых справедливы условия:

- $e_i - b_i > L$, где L – заданная величина, определяющая минимальную возможную длительность участка;
- $e_j – b_j < L$ для всех $j<i$ или $b_i – e_{i-1} >L$

если $e_i – b_i > L$, $e_{i-1} – b_{i-1} < L$ и $e_{i-2} – b_{i-2} > L$, то в качестве границ данного участка выбираются $b_{i-2}$ и $e_i$.

Схема включения модуля сегментации представлена на рис. 4.

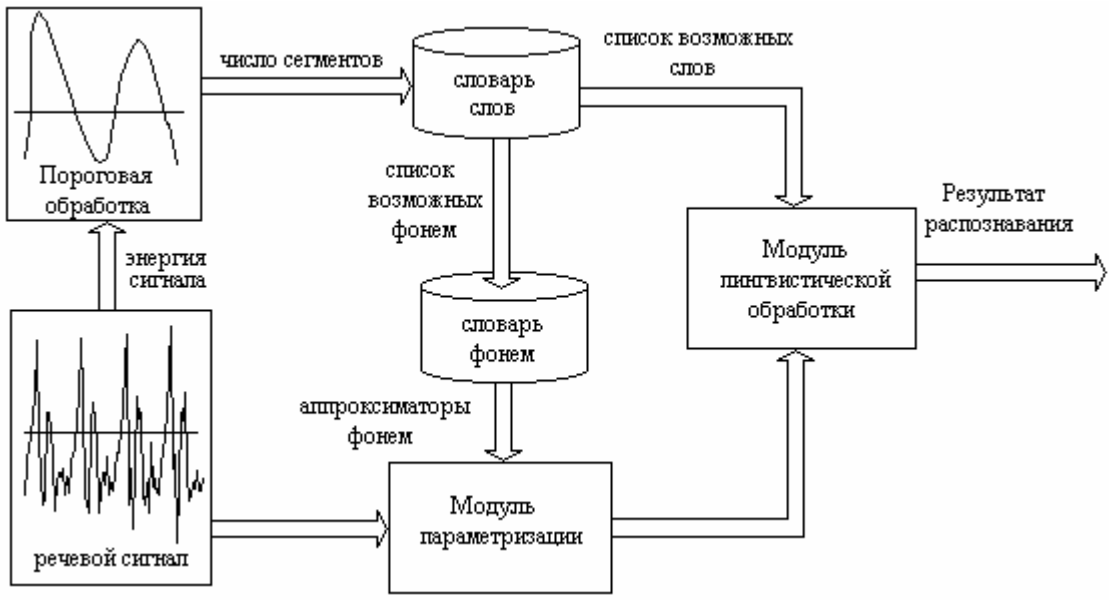


Рис. 4 – Схема включения модуля пороговой обработки в систему распознавания речевых команд

Таким образом, в схему распознавания включается дополнительный модуль сегментации, а лингвистическая оценка производится в виде последовательности оценок вокализованных частей высказывания. Полученные в результате экспериментов параметры составляют: N=60 мс, P=1800 и L=70 мс. Точность сегментации, достигнутая при использовании приведенного алгоритма, составляет 95%. Эта цифра может быть повышена за счет введения альтернативных способов сегментации некоторых слов.

В качестве приложения разработанных алгоритмов была выбрана система нейросетевого прогнозирования с элементами речевого управления. Для работы с этой системой был сформирован набор из 60 речевых команд. Точность распознавания на множестве команд составила ~ 90%.

### *Выводы*

Эксперименты моделирования показали, что описанный метод может быть использован в качестве базового в системе автоматического распознавания речи. Основными его недостатками следует считать невозможность распознавания невокализованных фонем, а также неустойчивость к смене диктора при распознавании согласных. Для преодоления этих проблем необходимо привлечение дополнительных процедур предварительной обработки входной информации. Вместе с тем, ряд традиционных для задачи распознавания речи вопросов, связанных с нелинейной временной структурой и сложностью определения границ речевых элементов, при рассмотренном подходе снимается. Кроме того, в сравнении с другими нейросетевыми методами распознавания речи предложенный алгоритм имеет то преимущество, что задача хранения образов распределяется на множество отдельных нейросетей. При этом снижается их размерность, сокращаются временные затраты на процесс обучения и возрастает разрешающая способность системы в целом. Добавление в систему новых фонем требует только формирования соответствующих нейросетевых аппроксиматоров, а словарь слов представляет собой текстовый файл с транскрипциями.

Дальнейшая работа может вестись в направлении поиска алгоритма автоматического сопоставления нейросетей с фонемами, что позволит формировать словарь фонем более рационально и снимет проблему работы с различными дикторами за счет автоматического введения дополнительных сетей.

### *Литература*

1. Методы автоматического распознавания речи: В 2-х книгах. Пер. с англ./ Под ред. У. Ли. – М.: Мир, 1983. – Кн. 1. 328 с.
2. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наукова думка, 1987. – 262 с.
3. Lee C-H., Paliwal. K. K. Automatic speech and speaker recognition: advanced topics. Boston: Kluwer, 1996 – 517 p.
4. Федяев О.И., Гладунов С.А. Нейросетевой интерпретатор речевых команд для управления программными системами. – Труды 7-й всероссийской конференции «Нейрокомпьютеры и их применение», НКП-2001, Москва, 14-16 февраля 2001 г./ Под редакцией А.И. Галушкина. М.: Институт проблем управления, 2001 – с. 298-301.
5. Рабинер Л.Р., Шафер Р.В. Цифровая обработка речевых сигналов: Пер. с англ. - М.: Радио и связь, 1981. - 495 с.
6. Потапова Р. К. Речевое управление роботом. – М.: Радио и связь, 1989. – 248 с.

Поступила в редакційну колегію 28.12.2002