ОЦЕНКА ПАРАМЕТРОВ МЕТОДА НЕЙРОСЕТЕВОЙ АППРОКСИМАЦИИ ФОНЕМ

Федяев О. И., Гладунов С.А.
Донецкий национальный технический университет
fedyaev@r5.dgtu.donetsk.ua, gladunov@ukr.net

Abstract

Fedyaev O.I., Gladunov S.A., Estimation of parameters in neural phoneme approximation method. *This article describes experiments and their results, conducted in order to choose optimal values of different parameters in a new speech recognition method. A main criterion of a choice is recognition accuracy. Whole set of parameter values should allow realizing a full-functional isolated words recognition system.*

Задача оценки параметров метода распознавания

В настоящее время не существует точного метода решения задачи распознавания звуковых образов [1], что связано с неоднозначностью акустического представления речи и отсутствием математического описания связи R между акустическим A и символьным S представлением речи. Существующие подходы являются недостаточно точными и строятся на основе субъективного синтеза структуры связи R. При этом возникает задача идентификации параметров разработанной модели связи R.

В настоящей статье рассматривается решение задачи параметрической идентификации предложенной в [2] модели R/, осуществляющей распознавание речи методом нейросетевой аппроксимации фонем. Структура модели R/ представлена на рис. 1. Блок контекстной обработки реализован на основе метода динамического программирования [3].

Рисунок 1 – Схема системы распознавания

Идентификация параметров формулируется в виде следующей нелинейной многомерной оптимизационной задачи:

$$\sum_{j=1}^{m}\delta(R(A_j),R'(A_j,T_{win},T_{\min},E,N,I,K_i,T))\xrightarrow[T_{win},T_{\min},E,N,I,K_i,T]{}\max,$$

где m – объем словаря; δ(x,y) – символ Кронекера; R(Aj)=Sj – преобразование j-го элемента словаря из акустического представления Aj в символьное Sj; R/ - предложенная модель; Twin, Tmin, E, N, I, Ki, T – параметры модели R/.

Оптимальные значения параметров модели R/ определялись посредством моделирования.

Определение параметров блока фрагментации

Реализация модуля сегментации требует определения 3 параметров:

- Tmin – минимальная длительность информативного/неинформативного участка. Это значение может быть определено исходя из минимальной длительности произнесения одной фонемы. Опыт ручной сегментации слов по их графическому представлению показывает, что рациональным значением является Tmin = 350 мс.
- Twin – размер скользящего окна, в рамках которого вычисляется энергия сигнала;
- E – пороговое значение энергии активного фрагмента;

Значение Tmin может быть определено исходя из минимальной длительности произнесения одной фонемы. Опыт ручной сегментации слов по их графическому представлению показывает, что рациональным значением является Tmin = 350 мс. Оптимальные значения Twin и E определялись путем моделирования. В качестве критерия был выбран показатель, определяемый количеством альтернативных способов сегментации различных вариантов произнесения одного и того же слова. Способ сегментации определялся количеством информативных сегментов.

При моделировании рассматривался набор из 60 слов, использованных в качестве команд речевого управления информационной системой нейросетевого анализа данных. Каждое слово было представлено 5 вариантами произнесения.

По результатам экспериментов построена поверхность, отображающая эмпирическую зависимость качества сегментации от размеров окна и значения порога (рис. 2). При этом каждому слову приписывалась основная сегментация, присущая наибольшему числу вариантов его произнесения. Качество определялось как процент слов с сегментацией, совпадающей с основной для данного слова. Значения порога Ротн на графике нормированы и вычисляются по формуле:

$$P_{отн} = \frac{E}{T_{win} * F * 2^{n-1}} * 100,$$

где Ротн – значение порога в процентах; F – частота дискретизации сигнала; n – число бит квантования сигнала. В описываемых экспериментах F=8000 и n=8.

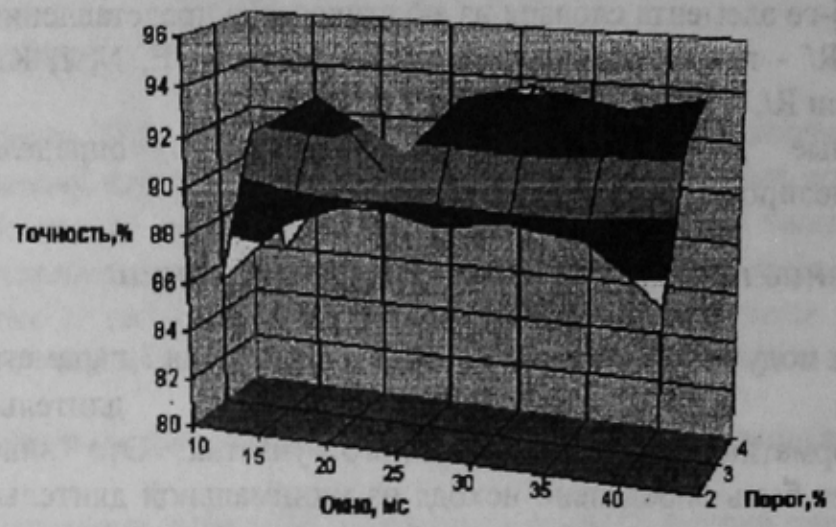

Рисунок 2 – Зависимость точности сегментации сигнала от размера окна и значения порога

Наибольшая точность сегментации (94,3%) достигается при длительности окна в 30 мс и значении порога 3%. Результаты получены при шаге длительности окна, равном 5 мс и шаге порога 1%. Точность алгоритма сегментации определяет верхнюю границу качества распознавания. Однако этот параметр может быть улучшен за счет введения двойной сегментации некоторых слов. Из 17 ошибок сегментации (5,7%), полученных в наиболее удачном эксперименте, неправильно сегментированы 12 слов, из них в 7 ошибка встречается только на одной реализации, а в 5 – на двух. Слово с 3 вариантами сегментации было единственным. Таким образом, добавив в словарь по дополнительному варианту сегментации 4 слов, можно увеличить точность алгоритма до 97%.

Выбор способа нормализации речевого сигнала

Значения сигнала $A_w(t)$, формирующие входной образ для распознавания методом нейросетевой аппроксимации фонем, должны быть предварительно нормированы. Это обусловлено двумя причинами:

- нестабильностью громкостной компоненты сигнала;
- ограниченностью нейросетевой функции активации диапазоном [0; 1].

В настоящей работе выбор процедуры нормирования производился по результатам моделирования. Рассматривались следующие способы:

– ортогональное проецирование обучающих пар на отрезок [0;1] (Norm1);
– ортогональное проецирование всего сигнала на отрезок [0;1] (Norm2);
– преобразование всех значений исходного сигнала по формуле

$$A'_w(t) = \frac{1}{1+e^{-A_w(t)}},$$

где $A_w(t)$ – текущее значение амплитуды сигнала; $A'_w(t)$ – нормированное значение (Norm3);

– нормировался всего сигнала по формуле (Norm4)

$$A'_w(t) = \frac{\sin(A_w(t))+1}{2};$$

– преобразование всего сигнала по формуле (Norm5)

$$A'_w(t) = \frac{arctg(A_w(t+1)-A_w(t))+\pi}{2\pi}.$$

Для оценки влияния каждого из способов нормализации был использован набор из 10 слов («икра», «нерпа», «экран», «сирена», «сатира», «раритет», «принтер», «сварка», «ветка», «отвар»), каждое из которых представлялось 10 вариантами произнесения. Все слова содержат два активных сегмента. Точность сегментирования на выбранном словаре составила 100%. В словарь фонем вошли звуки «а», «р», «и», «э», «в». Нейросетевая аппроксимация фонем проводилась с использованием 3-слойных сетей с 20 входами и распределением нейронов по слоям 20-10-1. Обучение осуществлялось на фонемах, выделенных из приведенных слов, что позволяет учесть влияние коартикуляции. Результаты качества распознавания, полученные каждым из способов нормализации, приведены на рис. 3.

Рисунок 3 – Зависимость качества распознавания от используемого способа нормирования сигнала

Эксперимент показал, что лучший результат (68%) достигается при использовании способа нормирования Norm1. Этот способ, несмотря на наибольшие вычислительные затраты на каждом шаге, дает результат, почти в 3 раза превосходящий лучший из остальных способов (25%).

Определение параметров нейросетевых аппроксиматоров

Для определения наиболее удачных параметров нейросетевых аппроксиматоров был использован словарь, описанный в 2. В ходе экспериментов нейросети обучались по фонемам, выделенным из слов словаря. Значения параметров выбирались в зависимости от качества распознавания.

Первым из рассматриваемых параметров было количество слоев нейросети N. Использовались сети с количеством слоев от 1 до 3 и одинаковым количеством входов. Рассматривать сети с большим количеством слоёв нецелесообразно [4]. Выходной слой должен состоять из 1 нейрона, что определяется решаемой задачей. Количество входов сети также зафиксируем. В этом случае однослойный вариант сети будет состоять из единственного нейрона с заранее определенным количеством входов. Двух- и трехслойные сети могут иметь различное количество нейронов в скрытых слоях, однако это количество выбиралось таким образом, чтобы число весовых коэффициентов в сетях было одинаково. Для повышения достоверности выводов было исследовано 3 группы сетей. Результаты моделирования представлены на рис. 4.

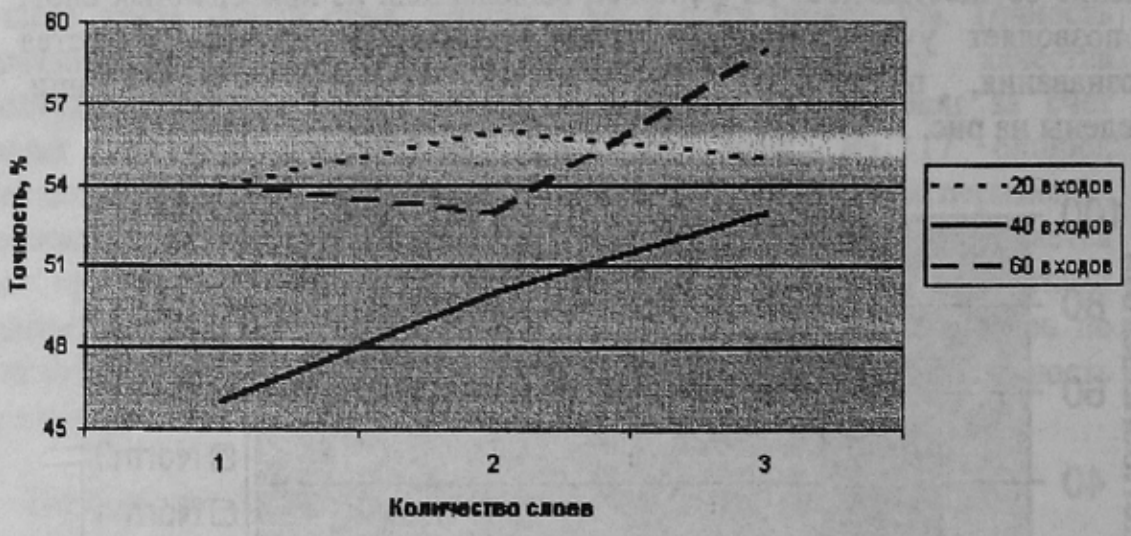

Рисунок 4 – Зависимость точности распознавания от количества слоев

Как показали результаты, наилучшее распознавание в 2 случаях из 3 достигается при использовании 3-слойной сети. Исключение составили сети на 20 входов, однако разница в этом случае составляет всего 1%.

Таким образом, более удачными для аппроксимации фонем можно считать 3-слойные сети.

Следующим параметром нейросетевых аппроксиматоров является число входов I. Для исследования зависимости от него точности распознавания использовалась та же модель, а также 3-слойные сети с распределением нейронов по слоям 2-2-1. Полученная зависимость показана на графике (рис. 5).

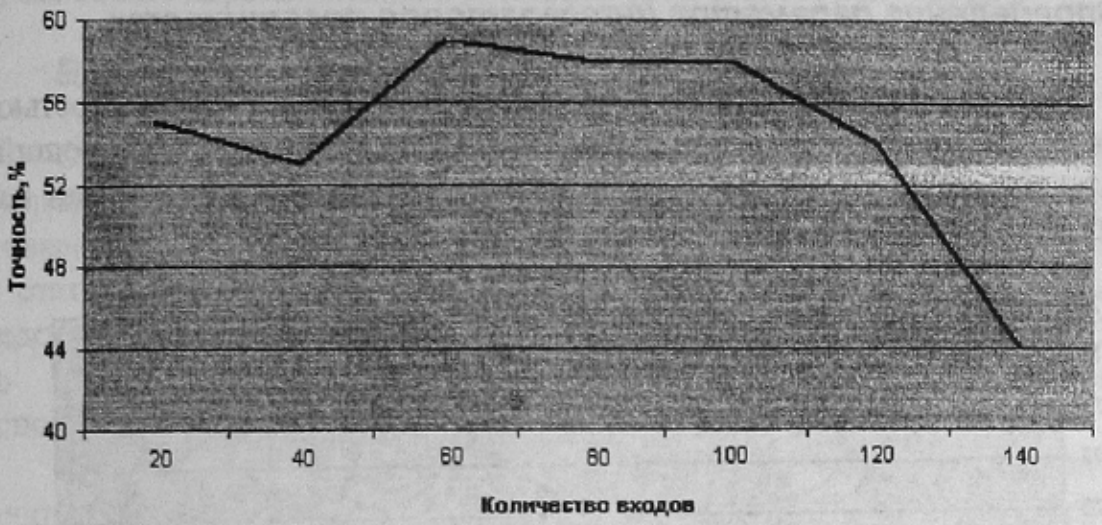

Рисунок 5 – Зависимость точности распознавания от размера входного окна

Наиболее высокая точность (59%) достигнута при использовании сетей с 60 входами, т.е. длительностью входного окна 7,5 мс. Результаты для 80- и 100-входовых нейросетей одинаковы и составляют 58%, т.е. почти не отличаются от максимального. Поэтому предпочтение было отдано модели с наименьшим числом параметров.

Последним неопределенным параметром нейросетевых аппроксиматоров остается распределение числа нейронов по слоям Ki. В табл. 1 приведены результаты моделирования распознавания при различных количествах нейронов скрытых слоев.

Таблица 1. Зависимость точности распознавания от количества нейронов в скрытых слоях нейросетевых аппроксиматоров

Размерность сети	Количество весов	Точность распознавания, %
60/10-10-1	710	62
60/10-20-1	820	43
60/20-10-1	1410	76
60/20-20-1	1620	63
60/30-10-1	2110	65
60/30-20-1	2420	48
60/30-30-1	2730	58

60/40-20-1	3220	50
60/50-10-1	3510	57
60/60-10-1	4210	55

Наиболее удачный результат (76%) получен при использовании сети 20-10-1 на 60 входов. Дальнейшее увеличение размеров слоев нейросети не приводит к росту точности распознавания.

Определение параметра интеграторов погрешности

Интеграторы погрешности характеризуются только длительностью времени накопления T. Экспериментальный график, отображающий зависимость точности распознавания речи от этого параметра показан на рис. 6.

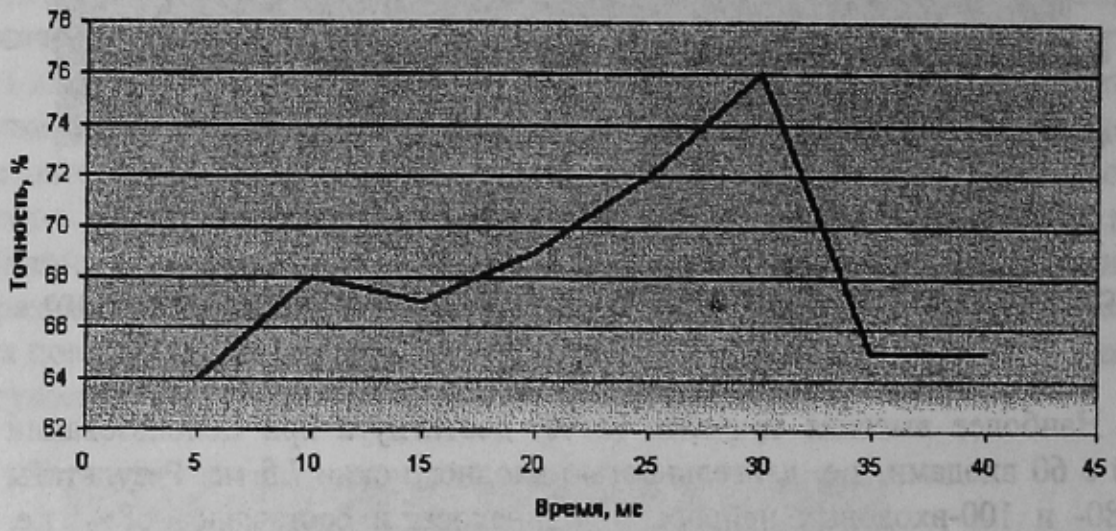

Рисунок 6 – Зависимость точности распознавания от длительности такта интегратора

На графике видно, что наиболее высокая точность распознавания (76%) была достигнута при значении T=30 мс. Это значение обратно пропорционально объему памяти и вычислительным затратам, требуемым на оценку сходства цепочек фонем методом динамического программирования. Таким образом, с точки зрения оптимизации вычислительного процесса, большие значения параметра T являются более предпочтительными. Однако на графике видно, что при росте T сверх 30 мс точность распознавания значительно снижается и поэтому дальнейшее увеличение рассматриваемого параметра нецелесообразно.

В предложенной схеме распознавания имеется ещё один фактор, связанный с использованием сегментации слов. Поскольку слово представляется в виде совокупности активных участков, на каждом из

которых вычисляется соответствующая степень сходства с цепочками словаря, возможно два подхода к оценке совокупной достоверности:

- достоверности цепочек суммируются в виде абсолютных значений;
- достоверности цепочек равномерно нормируются, а затем суммируются.

Оба подхода были промоделированы на словаре из 60 слов, каждое из которых было представлено 5 вариантами произнесения. В первом случае точность распознавания составила 89%, а во втором – 81%.

Выводы

Описанные эксперименты позволили реализовать речевой интерпретатор команд на примере системы нейросетевого анализа данных. Словарь системы составил 60 слов. Качество распознавания определялось на статическом тестовом множестве, в котором каждая команда была представлена 5 реализациями. Точность распознавания составила 91,3% что является приемлемым показателем в реальных системах распознавания.

Вместе с тем, реализация метода показало, что процесс распознавания требует значительного времени (3-5 секунд на команду, в зависимости от быстродействия компьютера), что приводит к снижению эффективности речевого взаимодействия. Поэтому необходимы исследования возможности ускорения работы метода за счет его аппаратной реализации с учетом возможности значительного распараллеливания нейросетевого алгоритма.

Литература

1. R. Chengalvarayan. Evaluation of front-end features and noise compensation methods for robust Mandarin speech recognition, Proc. Eurospeech, 2001 – pp. 583-598.
2. Гладунов С. А., Федяев О. И. Нейросетевой метод фонетической сегментации речевого сигнала. – Науч. тр. Донецкого гос. тех. университета. Серия: Проблемы моделирования и автоматизации проектирования динамических систем, вып. 52, 2002 – с. 125-130.
3. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов. Киев: Наукова думка, 1987. – 262 с.
4. Нейроинформатика / А.Н.Горбань, В.Л.Дунин-Барковский, А.Н.Кирдин и др. - Новосибирск: Наука. Сибирское предприятие РАН, 1998. – 296 с.

Дата надходження до редколегії: 3.12.2003 р.