Распознавание речевых слов с помощью искусственных нейросетей

О.И. Федяев, С.А. Гладунов
Кафедра ПМИ, ДонГТУ
E-mail: fedyaev@r5.dgtu.donetsk.ua

Abstract

Fedyaev O.,Gladunov S. Human speech recognition using artificial neuronets. *In this work it was considered a solution of problem of human speech recognition using artificial neuronet (a back propagation model). Entrance sets formed by spectral analysis. Analysis of recognition was conducted with methods of simulation. The best result reached is 92 %.*

Введение

Речь традиционно считается самой распространенной формой человеческого общения. Речевое общение достигло такой степени совершенства, что в произносимых словах можно передать гораздо больше информации, чем в печатном тексте. Поэтому решение задачи автоматического распознавания речи (automated speech recognition - ASR) открывает широкий спектр возможных применений этой технологии, начиная с автоматизации ввода текста под диктовку и кончая созданием систем безопасности и контроля доступа на основе технологии речевой подписи. В настоящее время выпущены на рынок следующие продукты ASR: пакет Voice 1.0 for Windows компании Kurzveil Applied Intelligence, Dragon Dictate компании Dragon System, Voice Type Dictation for OS/2 and Windows фирмы IBM, Listen 2.0 for Windows компании Vorbex Voice Systems.

Тем не менее, на сегодняшний день эта задача не была решена качественно. Большинство современных продуктов автоматического распознавания речи способны работать только с дискретным речевым вводом, предполагающим четкое произнесение слов с достаточной паузой между ними. Не решена проблема изменения гармонического состава голоса говорящего от сеанса к сеансу. Чтобы система речевого ввода "схватывала на лету" произносимую речь, она должна в реальном времени выделять, преобразовывать в цифровую форму и распознавать поток произносимых слов.

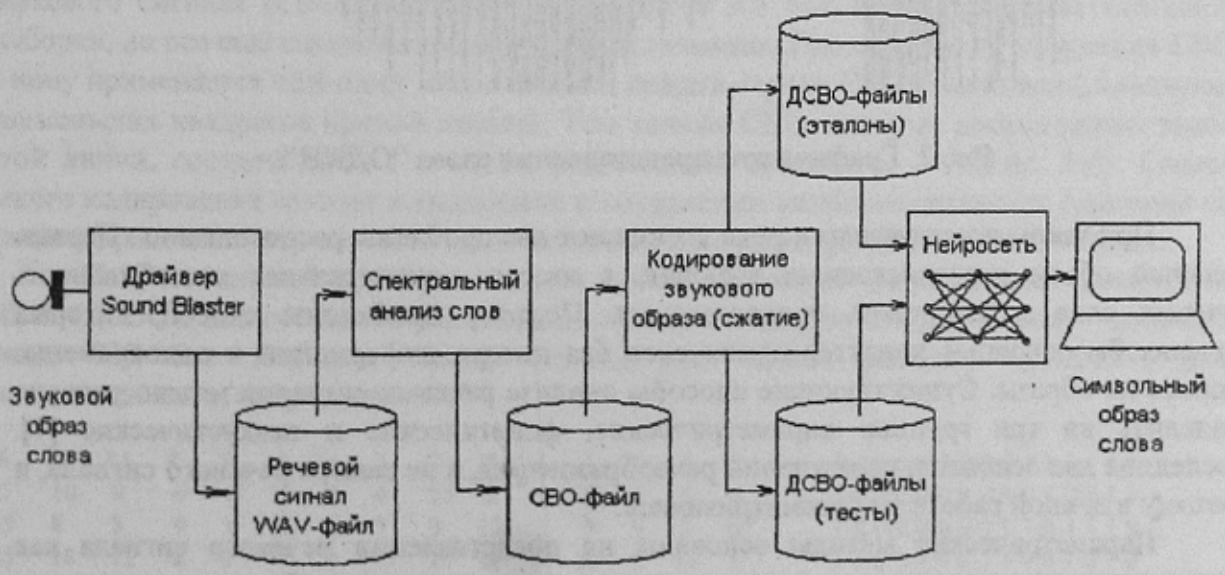

Рис. 1. Схема распознавания слов.

1. Структурная схема распознавания

Одним из многообещающих способов улучшения качества систем распознавания можно считать использование искусственных нейронных сетей. В чем новизна нейросетевого подхода к проблеме автоматического распознавания речи? Анализ работ в этом направлении показывает, что с помощью нейронных сетей удается более качественно проводить акустико-фонетический анализ. Поэтому цель данной работы оценить достоинства и недостатки нейросетей как инструмента для распознавания речевых слов.

Исследование процесса распознавания проводилось по функциональной схеме (рис 1). Звуковой сигнал с микрофона поступает в Sound Blaster компьютера, который формирует соответствующий WAV-файл. Спектральный анализ звукового файла дает спектральный временной образ принятого сигнала (СВО). В спектральном временном образе отражены все особенности речевого сигнала. Линейная аппроксимация СВО позволяет существенно сжать входной образ и, в то же время, сохранить характерные признаки распознаваемого сигнала. В результате получается двоичный спектральный временной образ (ДСВО), который пополняет каталог эталонов или поступает на нейросеть для распознавания.

2. Формирование цифровых образов речевых слов

Изменение акустического давления с частотой 16-16000 Гц называется звуком. Современные компьютеры, оснащенные Sound Blaster, позволяют записывать звук в цифровом виде. При этом акустическое давление считывается через равные промежутки времени.

Представляемый в цифровом виде звуковой сигнал должен иметь достаточную частоту дискретизации, чтобы сохранять все свойства звукового образа, но не слишком большую, чтобы не нести избытка информации. Обычно рекомендуется частота дискретизации 8 кГц.

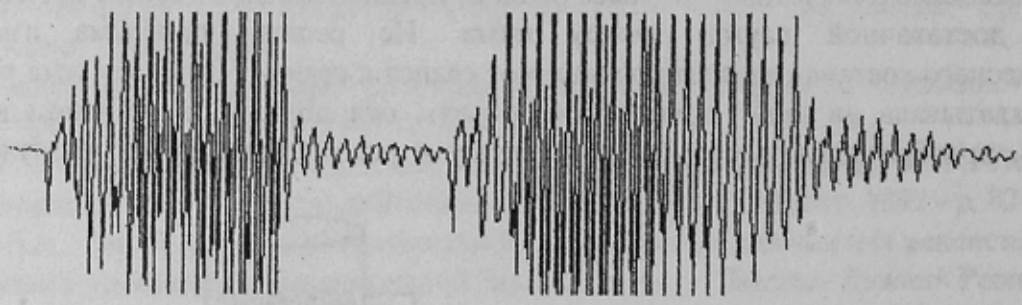

Рис 2. Графическое представления слова “ОДИН”.

При таком представлении речи возникают две проблемы распознавания. Первая- большой объем распознаваемых выборок, а вторая - значительная нестабильность речевых слов в цифровом представлении. Поэтому необходим способ, который выделял бы основные характеристики слов без потери информации и одновременно сжимал их образы. Существующие способы анализа речевых сигналов можно условно разделить на три группы: параметрические, фонетические и неакустические [1]. Последние два основаны на изучении речеобразования, а не самого речевого сигнала, и поэтому в данной работе не рассматривались.

Параметрические методы основаны на представлении речевого сигнала как реализации некоторого процесса во времени и выделения каких-либо параметров этого

процесса, чаще всего связанных тем или иным образом с его спектральными характеристиками. К этим методам относятся:

- спектрально-полосные;
- ортогональные;
- корреляционные;
- метод непосредственного вычисления спектра с помощью быстрого преобразования Фурье;
- методы, связанные с выделение мгновенной частоты переходов через нуль клиппированного речевого сигнала (ρ-параметров);
- временные методы, основанные на анализе распределения длительностей интервалов между переходами через нуль или экстремумами речевого сигнала;
- методы, использующие нелинейные преобразования и фазовые соотношения речевого сигнала.

В настоящей работе использовался метод разложения звукового сигнала в ряд Фурье с последующим выделением максимумов амплитуд гармоник с сохранением их местоположений .

При использовании метода Фурье слово разбивается на равные промежутки длиной 10 мс (80 отсчетов), каждый из которых рассматривается как заданная по времени табличная функция $f(t)$. После этого для каждого участка вычисляется набор амплитуд гармоник A_k с частотами 100, 200, 300, ... 1500 Гц по формулам:

$$A_k = \sqrt{a_k^2 + b_k^2} ,$$

где $$a_k = \frac{1}{N}\sum_{t=0}^{2N-1} f(t)\cdot\cos\frac{\pi\cdot k\cdot t}{N} ;$$

$$b_k = \frac{1}{N}\sum_{t=0}^{2N-1} f(t)\cdot\cos\frac{\pi\cdot k\cdot t}{N} ;$$

$$a_0 = \frac{1}{N}\sum_{t=0}^{2N-1} f(t); \quad b_0 = 0 .$$

Звуковой образ при таком разложении можно представить в виде двумерной таблицы, где по вертикали откладывается время, а по горизонтали - номер гармоники (рис. 3,а).

Полученная таблица представляет собой спектральный временной образ звукового сигнала (СВО). Его объем примерно в 3,3 раза меньше объема исходной выборки, но все ещё слишком велик для распознавания. Поэтому после получения СВО к нему применяется ещё один метод сжатия: каждая строка СВО сглаживается методом наименьших квадратов прямой линией. Тем точкам СВО, которые расположены выше этой линии, соответствует 1 в новом разложении, остальным - 0 (Рис. 3,б). Смысл такого кодирования состоит в выделении и сохранении наиболее активных гармоник на каждом временном отрезке. Эти пики активности и характеризуют звуковой сигнал.

Полученная выборка представляет собой двоичный спектральный временной образ (ДСВО). Его объем в 8 раз меньше объема исходного СВО и обычно занимает около 700 бит. Выборки такого объема уже можно реально применять для распознавания на нейросети.

36	14	11	8	9	9	6	5	7	7	6	7	5	5	0	1000000001111110
27	10	9	6	6	6	2	4	11	2	5	2	2	5	0	1000000010101111
35	8	5	9	6	10	4	7	9	12	2	5	6	2	0	1000000011011110
25	18	12	6	3	7	3	17	17	4	5	10	6	11	1	1100000110010101
40	26	18	13	12	18	15	16	13	15	11	3	9	8	0	1100010101101101
22	20	7	4	6	4	6	25	23	9	3	14	10	23	2	1100000110010101
36	21	7	7	6	9	4	21	24	15	10	4	9	5	1	1100000111110100

11	9	6	9	5	9	10	25	13	1	4	4	6	7	1	1000011111000010
26	21	15	13	7	14	17	9	11	15	4	7	8	6	1	1100001011001100
2	18	11	6	8	7	13	3	9	4	3	2	4	3	0	0110001010001100
40	28	16	10	8	15	12	7	1	3	2	3	4	2	0	1100011000011111
49	88	33	7	11	16	10	10	6	4	5	3	5	3	0	1100000000111111
46	132	25	24	21	16	24	7	12	4	11	5	8	5	0	0100000000111111
81	84	37	20	60	10	21	5	9	24	6	11	5	7	0	1100100001011111
82	57	49	11	24	15	12	5	10	10	4	1	5	5	0	1110000001011111
59	18	22	4	4	1	3	2	1	1	1	1	1	1	0	1010000000111111
31	28	5	2	3	2	2	1	2	1	1	1	1	1	0	1100000000111111
32	25	14	7	5	3	3	3	2	2	2	2	2	1	0	1100000000111111
43	14	17	6	5	3	3	2	2	1	2	1	1	1	0	1010000000111111
40	22	9	2	1	0	0	0	1	0	0	0	0	0	0	1100000000111111
34	26	18	6	5	5	3	3	3	2	2	2	2	2	0	1110000000111111
29	21	15	8	4	4	3	2	3	2	2	2	2	2	0	1110000000111111
57	39	14	23	4	11	2	4	5	4	3	2	3	0	0	1101000000111111
88	52	61	5	7	15	5	8	6	3	1	3	1	2	0	1110000000111111
69	51	70	19	21	8	19	8	3	10	5	5	5	4	0	1110000001111111
60	71	91	27	15	22	27	4	8	3	4	6	5	3	0	0110000000001111
64	75	78	28	3	12	14	10	6	7	8	7	2	2	0	1110000000111111
51	68	100	32	9	13	16	18	15	11	2	10	9	2	0	0110000000001111
46	65	109	23	25	12	11	21	2	9	3	2	5	3	0	0110000000001111
50	68	96	34	21	13	8	23	22	9	3	7	6	2	0	0110000010011111
46	91	74	34	28	17	10	13	7	9	3	3	5	4	0	0110000000001111
68	73	31	47	61	34	35	10	23	18	8	6	3	11	1	1100101000000011
58	75	24	53	51	30	40	13	7	12	15	1	15	6	1	0101101000010111
55	65	53	78	31	12	23	11	13	6	9	9	6	7	0	0111000000001111
73	71	31	71	8	15	24	9	10	12	7	3	7	4	0	1101000001111111
104	51	24	40	13	17	23	8	11	6	8	4	6	8	0	1100000000111111
111	38	30	27	3	13	10	8	10	6	3	3	6	4	0	1000000000111111
71	54	21	18	14	4	7	0	5	7	0	5	4	3	0	1100000001011111
70	35	44	30	13	7	6	6	3	3	3	4	5	3	0	1010000000111111
63	33	36	12	9	3	5	2	3	1	2	1	4	2	0	1110000000111111
52	39	31	12	3	6	6	4	2	3	3	3	1	2	0	1110000000111111
32	28	38	12	12	9	8	7	6	5	5	5	3	3	0	1110000000111111
26	34	16	8	7	5	5	4	4	3	3	2	3	2	0	1100000000111111
37	33	10	2	1	1	0	0	0	0	0	0	0	0	0	1100000000111111
24	12	9	5	2	2	1	1	1	1	1	0	1	1	0	1100000000111111
14	4	1	0	1	0	0	1	0	0	0	0	0	0	0	1000000000111111

а) б)

Рис 3. Дискретное представление слова “ОДИН”:
а) - спектральный временной образ; б) - двоичный спектральный временной образ.

3. Построение нейроалгоритма и обучение нейросети

Для распознавания звукового образа применялась многослойная однородная нейронная сеть обратного распространения. Нейроалгоритм распознавания речи на сети характеризуется следующими параметрами:

- **входной сигнал сети** - вектор X , представляющий собой развертку ДСВО;
- **выходной сигнал** - вектор Y , формируемый сетью как код слова;
- **желаемый выходной сигнал** - правильный код распознаваемого слова;
- **структура нейросети** - трехслойная с полными последовательными связями; модель искусственного нейрона включает сигмоидальную функцию активации $f(g) = 1/(1+e^{-g})$;

- **функция ошибки нейросети** - отклонение реального выходного сигнала от желаемого;
- **критерий качества обучения** - минимум ошибки распознавания на всем обучающем множестве;
- **весовые коэффициенты** - матрицы вещественных чисел, принадлежащих выходному и скрытым слоям сети.

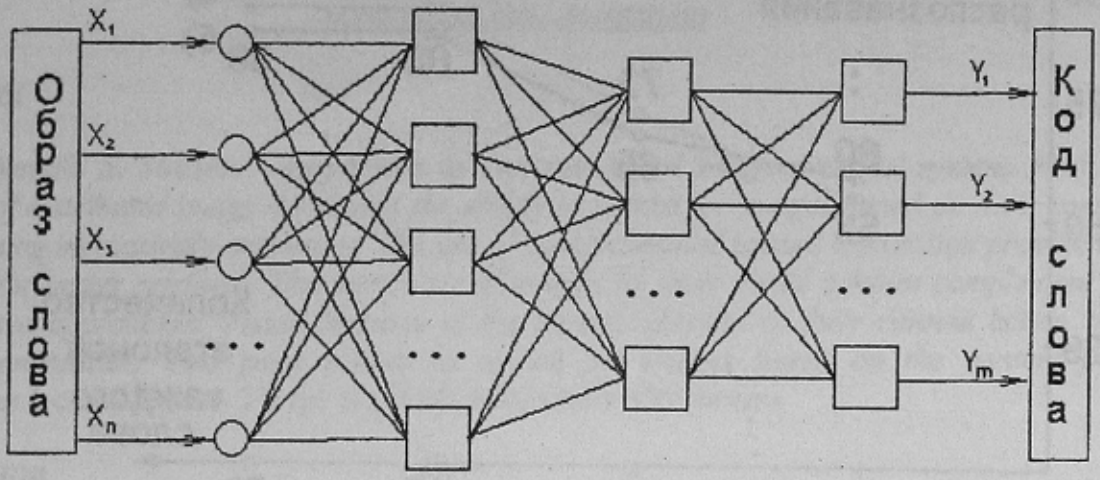

Рис 4. Нейросеть обратного распространения.

Нейронная сеть обучалась по стратегии "обучение с учителем" с помощью процедуры обратного распространения ошибки (Back Proposition) [2].

4. Результаты распознавания речи с помощью нейросети

Эффективность нейросетевого подхода оценивалась на примере распознавания пяти слов - "один", "два" , "три" , "четыре", "пять". Для обучения сети было построено обучающее множество, включающее от пяти до двадцати вариантов произношения каждого слова.

Рациональные размеры сети определялись экспериментально. Среди рассмотренных вариантов наилучшим образом зарекомендовала себя трехслойная сеть с распределением нейронов по слоям 8-7-6. Число входов выбиралось исходя из максимально возможного размера ДСВО, поэтому размер входного вектора составил 840 бит. Начальные приближения весовых коэффициентов задавались случайным образом из диапазона [-1 , 1]. Параметр скорости обучения не изменялся в процессе настройки сети и равнялся 0,2. Сеть считалась обученной, когда максимальное расхождение между реальным и желаемым выходными сигналами не превышало заданной величины $\xi = 0{,}2$. Для тестирования обученной нейросети применялось множество из семидесяти пяти образов указанных слов, не входящих в обучающее множество. При этом качество распознавания оценивалось на сети, обученной на 5,10,15 и 20 вариантах произношения каждого слова. Результаты распознавания приведены на рис. 5.

По оси абцисс отложено количество обучающих примеров, сформированных для каждого слова. Наилучшее распознавание составило 92%. Эксперименты показали, что увеличение числа эталонов свыше 15 не приводит к улучшению распознавания. Поэтому в каталоге образцов, формируемом для реальной системы ASR, достаточно хранить около 15 примеров произношения каждого слова.

Исследовалось влияние начальных значений весовых коэффициентов сети на скорость и качество обучения. В среднем нейронная сеть на компьютере типа 486 DX с частотой 100 MHz обучалась в течение 10 минут. Экспериментально было установлено,

если обучение начинать не со случайных значений весов, а дообучать ранее настроенную сеть на новое множество эталонов, то скорость обучения и качество распознавания улучшаются, что видно из сравнения графиков а) и б) на рис.5. Это, по-видимому, обусловлено большим значением параметра расхождения ξ. Однако уменьшение данного параметра существенно увеличивает время обучения.

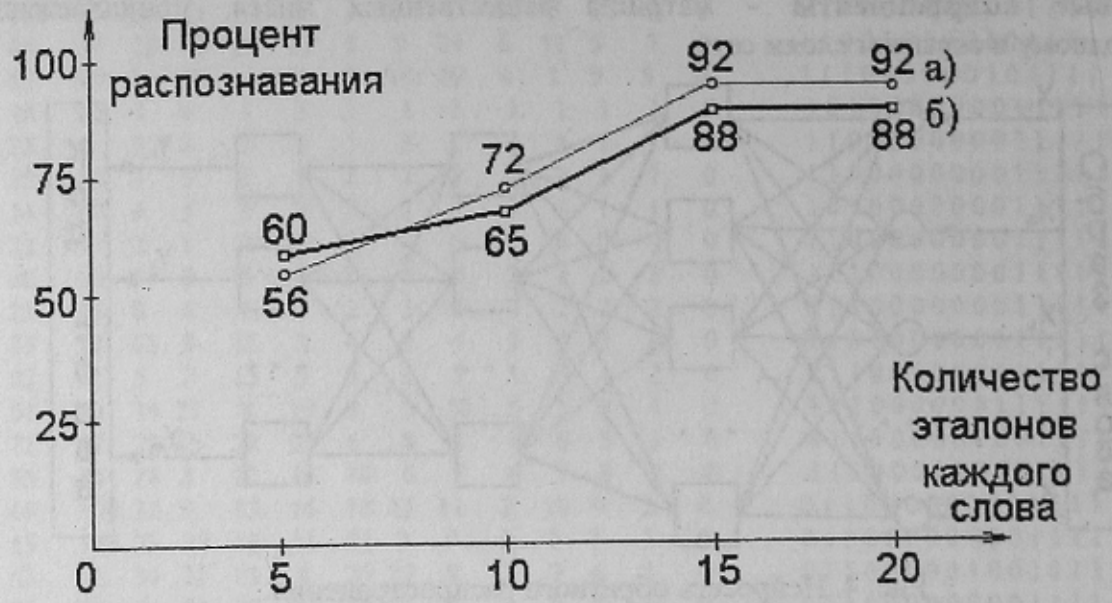

Рис. 5. Зависимость процента распознавания от числа эталонов слов:
а) сеть с дообучением; б) обучение со случайных начальных значений весов.

Заключение

Анализ экспериментов показал правомерность нейросетевого подхода к решению задачи автоматического распознавания речи. Его достоинствами можно считать высокую скорость распознавания, а также слабую чувствительность к изменениям тембра и громкости произношения. Однако для нейронной сети важно построение хорошего и компактного обучающего множества, адекватного реальной речи, что требует решения вопросов нахождения начала каждого слова, нормализации его по времени и сжатия речевого образа с сохранением характерных признаков.

Несмотря на трудности в формализации и решении указанных вопросов, повышение интеллектуализации общения пользователя с ЭВМ невозможно без разработки средств речевого ввода-вывода информации, поэтому выполненные исследования будут в дальнейшем использованы при разработке одного из видов системы ASR - интерпретатора команд для речевого управления различными приложениями (command-and-control).

Литература

1. Плотников В.Н. Речевой диалог в системах управления. - М.: Машиностроение, 1988. - 224 с.
2. Уоссермен Ф. Нейрокомпьютерная техника. Теория и практика. -М.: Мир, 1992. - 240с.