МНОГОУРОВНЕВАЯ НЕЙРОСЕТЕВАЯ СТРУКТУРА РАСПОЗНАВАНИЯ РЕЧЕВЫХ СЛОВ ПО НИЗКОЧАСТОТНЫМ ГАРМОНИКАМ

Федяев О.И., Гладунов С.А.

Кафедра ПМИ, ДонНТУ
fedyaev@r5.dgtu.donetsk.ua, gladunov@ukr.net

Abstract

Fedyaev O., Gladunov S. A neural structure with many layers to recognize of speech words by low-frequency harmonics. *In this work it was proposed a method of an isolated words recognition based on a decomposition of spectral pattern. It was designed a two-layer scheme of recognition based on an integral estimation of belonging harmonics of words and it's realization in a neural network basis. There are results described of a recognition process modeling.*

Введение

Наиболее привычным и естественным для человека является речевое общение, поэтому создание речевого пользовательского интерфейса компьютерных информационных систем представляет собой актуальную задачу. Несмотря на множество подходов к её решению, по-прежнему остается проблемой построение удобного для распознавания образа акустического сигнала.

В настоящей работе рассмотрены вопросы, связанные с разработкой нейросетевых структур и алгоритмов автоматического распознавания изолированных слов с целью создания интерпретатора речевых команд для эффективного управления информационными системами.

1. Значимость отдельных гармоник в структуре распознаваемых слов

Основной проблемой при распознавании речи является выбор компактного и информативного описания речевого сигнала, при котором существенно понижалась бы размерность образа слова и, при этом, сохранялись основные информативные признаки, позволяющие отличить одно слово от другого. Наиболее распространенным методом формирования цифрового представления является спектральный анализ, основанный на дискретном преобразовании Фурье. Он позволяет сгладить влияние случайной компоненты сигнала, достаточно устойчив к изменениям интенсивности (громкости) произнесения, но формируемый образ имеет большую размерность и неудобен для распознавания [1].

Анализ прямого и обратного преобразования слов показал, что для качественного распознавания нет необходимости использовать все гармоники. Определение минимально необходимого набора информативных спектральных составляющих позволяет существенно уменьшить образ речевого сигнала без ухудшения качества распознавания. Обратное преобразование исходного слова, хорошо воспринимаемое человеком на слух, получается по первым пяти низкочастотным гармоникам (на частотах 100, 200, … , 500 Гц). Это послужило

основанием для сокращения размерности распознаваемых образов за счёт использования только первых пяти низкочастотных гармоник.

Сокращенный спектральный временной образ (СВО) является адекватным и более компактным представлением исходного слова [2], но по-прежнему достаточно сложен для распознавания, что особенно проявляется при увеличении объёма словаря. Для дальнейшего упрощения задачи распознавания в работе предложен способ, который основан на декомпозиции входного образа к виду набора независимых подобразов, представляющих собой отдельные низкочастотные гармоники слова. Автономное рассмотрение отдельных гармоник одного слова возможно в связи со сходством их структур (рис. 1).

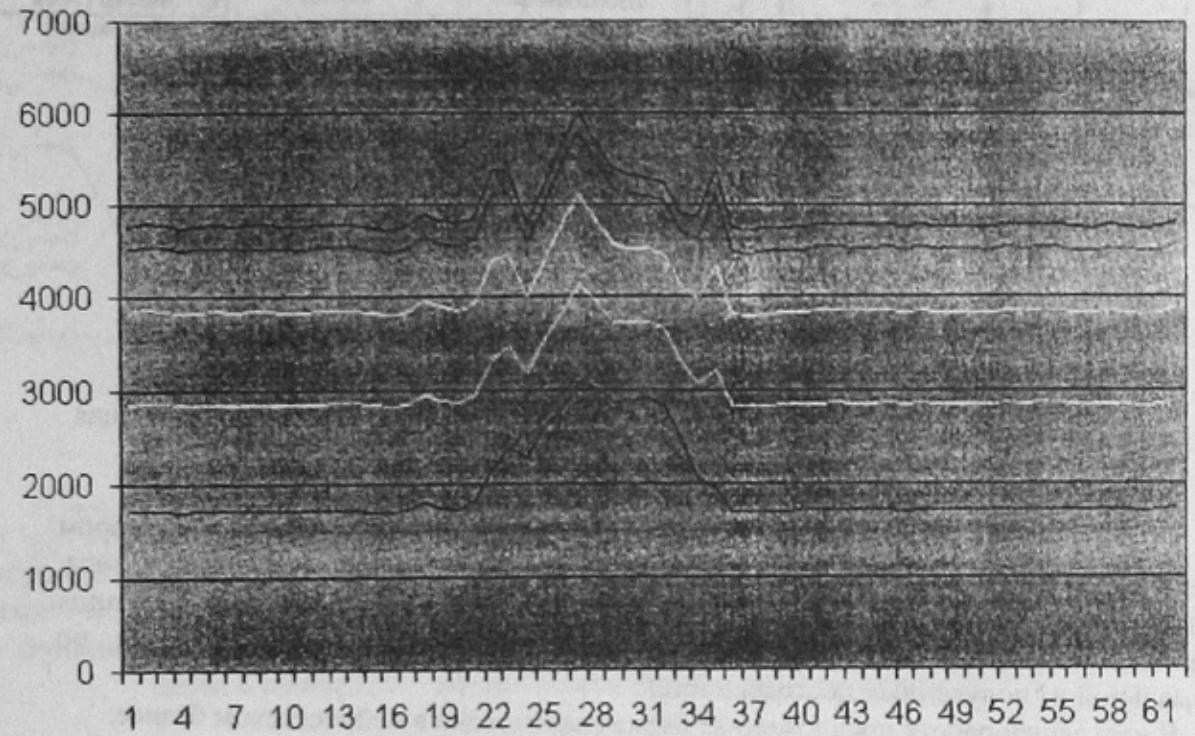

Рисунок 1 – Графическое представление первых пяти гармоник одного слова

Такой подход позволил снизить размерность распознаваемого образа до 40-60 чисел. В результате получается набор более простых подзадач распознавания, которые могут решаться независимо и параллельно.

2. Двухуровневая схема распознавания на основе декомпозиции спектрального образа слова

Распознавание слова по одной гармонике возможно лишь в идеальном случае. Реальная структура отдельных гармоник формируется под влиянием различных факторов, индивидуальных для конкретного произнесения слова, а также внешних помех. Кроме того, вклад отдельных гармоник в общую структуру слова неодинаков и, в частности, зависит от тембра голоса говорившего. В этой связи, наиболее надёжным видится решение, при котором на первом этапе в распознавании независимо друг от друга используются несколько гармоник, а затем частные результаты обобщаются, и на

их основе формируется окончательный вывод о распознаваемом слове. Именно такой многоуровневый способ распознавания схематически представлен на рис. 2.

Первый уровень распознавания предназначен для определения степени принадлежности гармоник словам словаря. Выводы о принадлежности делаются экспертными системами независимо друг от друга. При этом структура исходного слова определяется порядковыми номерами используемых гармоник и косвенно учитывается данной схемой.

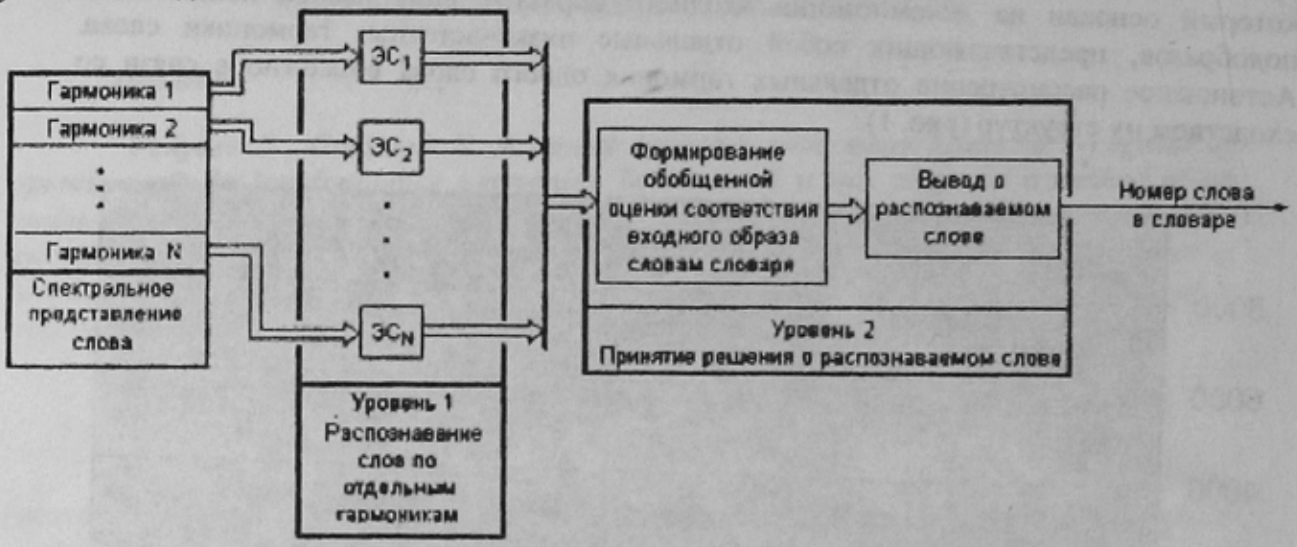

Рисунок 2 – Двухуровневая схема распознавания: $ЭС_i$ –экспертная система, делающая вывод о распознаваемом слове по i-й гармонике.

Полученные результаты первого уровня распознавания обобщаются на втором уровне в набор достоверностей соответствия исходного слова каждому из слов словаря. Обобщённые оценки носят более объективный характер, чем результаты распознавания отдельными экспертными системами. Окончательный выбор слова делается по максимальной из полученных достоверностей.

В настоящей работе предложенная схема реализована в нейросетевом базисе.

3. Иерархическая нейросетевая структура распознавания слов по отдельным гармоникам

Использование искусственных нейросетей для распознавания речи обусловлено их способностью к разделению сложных многомерных образов, устойчивостью к незначительным изменениям входного сигнала, а также возможностью динамического дообучения на новых примерах.

Для реализации элементов экспертной оценки принадлежности гармоник словам словаря (первый уровень распознавания на рис. 2) были выбраны однородные сети со следующим нейроалгоритмом:

- входной образ – распознаваемая гармоника;
- выходной сигнал – функция принадлежности гармоники словам словаря;
- желаемый выходной сигнал – вектор размерности словаря из нулей и одной единицы, соответствующей произнесённому слову;

- структура нейросети – трёхслойная с полными последовательными связями [2]; модель искусственного нейрона использует сигмоидальную функцию активации $f(g) = 1/(1+e^{-g})$;
- функция ошибки – отклонение реального выхода от желаемого;
- критерий качества обучения – минимум ошибки по всему обучающему множеству;
- обучение – при помощи алгоритма обратного распространения ошибки.

Результаты первого уровня распознавания передаются на второй уровень, схема реализации которого в нейросетевом базисе представлена на рис. 3.

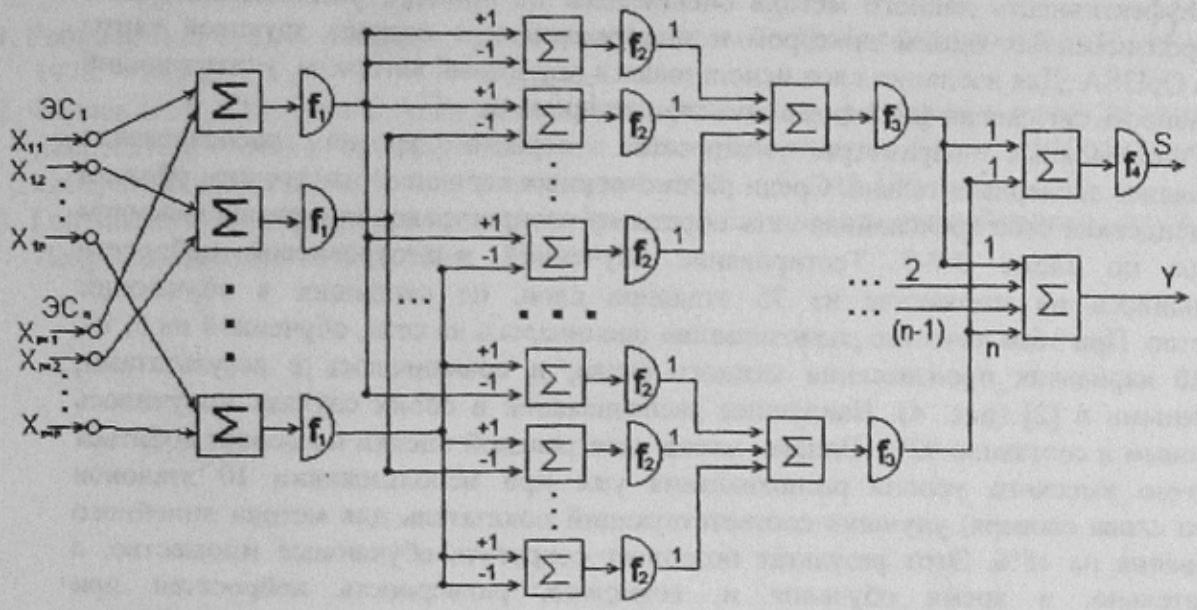

Рисунок 3 – Реализация второго уровня распознавания в нейросетевом базисе: X_{ij} – j-й выход i-й нейросетевой экспертной системы первого уровня распознавания; f_1, f_2, f_3, f_4 – функции активации соответствующих слоёв нейросети; Y – номер распознанного слова в словаре; S – сигнал о невозможности распознавания.

Как видно из рисунка, сеть второго уровня распознавания состоит из 4 слоёв. Первый слой выполняет интеграцию степеней соответствия распознаваемого слова каждому из слов словаря, полученных по отдельным гармоникам. Выход этого слоя можно интерпретировать как нечёткую функцию принадлежности слова. На этом уровне использовалась пороговая функция активации: $f_1(g)=\{1$, если $g \geq p$; 0, если $g<p\}$, где p – некоторое пороговое значение, зависящее от числа используемых гармоник m: $p=\alpha m$; α - коэффициент, получаемый эмпирически и характеризующий необходимый для распознавания уровень соответствия распознаваемого слова эталонному. Второй и третий слои предназначены для определения максимального уровня соответствия [3]. В этих слоях были использованы функции активации: $f_2(g)=\{1$, если $g \geq 0$; 0, если $g<0\}$ на втором слое и $f_3(g)=\{1$, если $g \geq n$; 0, если $g<n\}$ на третьем слое. Здесь n – количество слов в словаре. Нейроны третьего слоя формируют сигнал 1, если соответствующий входной элемент второго слоя максимален, или 0 в противном случае. На четвёртом слое определяется номер слова в словаре, соответствующего распознаваемому, а также вырабатывается сигнал S о невозможности распознавания. Здесь $f_4(g)=\{1$, если $g>1$; 0, если $g \leq 1\}$ – функция активации нейрона, делающего вывод о неспособности распознать входное слово с помощью данной нейросети. Подобная ситуация возникает в тех

случаях, когда ни одно входное слово не соответствует в достаточной степени ни одному из слов словаря, или наоборот – соответствует сразу нескольким. Нейрон, возвращающий номер слова в словаре (на схеме сигнал Y), имеет линейную функцию активации, поэтому на схеме она не отображена.

4. Эффективность распознавания речи методом интегральной оценки принадлежности гармоник словам

Эффективность данного метода оценивалась на примере распознавания пяти слов, произнесённых одним диктором и оцифрованных с помощь звуковой карты Yamaha Opl3SA. Для изоляции слов использовался пороговый алгоритм, учитывающий интенсивность сигнала на фиксированном отрезке времени.

Рациональные параметры нейросетей первого уровня распознавания определялись экспериментально. Среди рассмотренных вариантов наилучшим образом зарекомендовала себя трёхслойная сеть обратного распространения с распределением нейронов по слоям 8-7-5. Тестирование обученной многоуровневой нейросети производилось на множестве из 75 эталонов слов, не входящих в обучающее множество. При этом качество распознавания оценивалось на сети, обученной на 5, 10, 15 и 20 вариантах произнесения каждого слова, и сравнивалось с результатами, полученными в [2] (рис. 4). Наилучшее распознавание в обоих случаях получилось одинаковым и составило 92%. Однако, метод интегральной оценки позволяет добиться достаточно высокого уровня распознавания уже при использовании 10 эталонов каждого слова словаря, улучшив соответствующий показатель для метода линейного сглаживания на 18%. Этот результат позволяет сократить обучающее множество, а следовательно, и время обучения и, возможно, размерность нейросетей при использовании интегрального метода по сравнению с методом линейного сглаживания. Подобное улучшение качества достигается за счёт использования более адекватного обучающего множества, а также за счёт уменьшения размерности пространства признаков для каждой нейросети первого уровня.

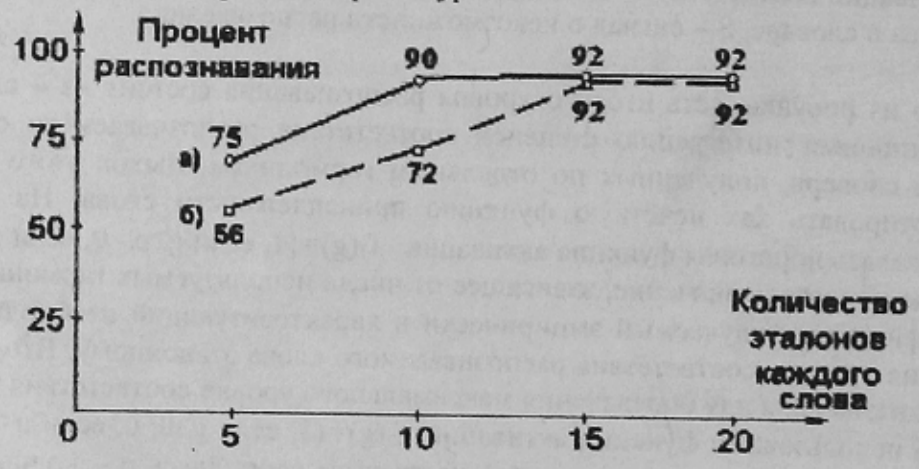

Рисунок 4 – Эффективность распознавания методами:
а) интегральной оценки; б) линейного сглаживания спектрального образа

Заключение

Предложенный в работе метод интегральной оценки позволяет сократить время обучения используемых нейросетей и повысить качество распознавания на меньшем

обучающем множестве. Предусмотрен механизм самоконтроля и определения невозможности распознавания. Удалось достичь лучшей устойчивости к изменениям громкости произнесения.

В дальнейшем предполагается продолжить анализ и усовершенствование метода с учётом увеличения объёма словаря и предварительной временной нормализации входного речевого образа.

Литература

1. Рабинер Л., Гоулд Б. Теория и применение цифровой обработки сигналов. – М.: Мир, 1978. - 848 с.
2. Федяев О.И., Гладунов С.А. Распознавание речевых слов с помощью искусственных нейросетей. – Науч. тр. Донецкого гос. тех. университета. Серия: Информатика, кибернетика и вычислительная техника, вып. 1999. – С. 145-150.
3. Галушкин А.И. О решении задач сортировки с использованием нейронных сетей // Нейрокомпьютер, № 3, 4, 1994. – С. 35 – 39.

Поступила в редакційну колегію 7.011.2001 р.