УДК 621.391

В.Т. Дмитриев, И.В. Баландин

ДИКТОРОНЕЗАВИСИМАЯ СИСТЕМА АВТОМАТИЧЕСКОГО ПОИСКА КЛЮЧЕВЫХ СЛОВ В ПОТОКЕ СЛИТНОЙ РЕЧИ, УСТОЙЧИВАЯ К АКУСТИЧЕСКИМ ШУМАМ

Представлена устойчивая к окружающим акустическим шумам диктороне-зависимая система автоматического поиска ключевых слов в потоке слитной речи. Основой предложенной системы является многослойная нейронная сеть, а для формирования первичных признаков речевого сигнала использовалось вейвлет-пакетное разложение. Рассмотрено деление системы на функциональные модули и приведено их описание, а также результаты работы системы. Показано, что применение предложенной системы позволит получить надежность успешного распознавания ключевых слов при отношении сигнал-шум равном 30 дБ не менее 90 %.

Введение. Ряд зарубежных компаний предлагают довольно широкий спектр систем распознавания речи - от голосового набора телефонного номера с настройкой на диктора до диктофонов с практически неограниченным словарем (до 100 тысяч слов), распознающих речь независимо от диктора или с адаптацией к диктору. В то же время анализ данных систем свидетельствует о высокой зависимости характеристик распознавания от типа конкретного диктора и уровня акустических шумов, что не позволяет получить надежность распознавания выше 70 – 80 % [1, 2].

Системы поиска ключевых слов находят широкое применение в устройствах управления голосом, системах голосового доступа, а также в устройствах защиты речевой информации. В связи с этим актуальной является задача распознавания ключевых слов в потоке слитной речи при наличии акустических помех.

При реализации таких устройств трудности возникают при выборе математических методов работы решающего устройства и первичных элементов речи, которые значительно влияют на надежность распознавания. Это связано с изменчивостью акустического образа, приписываемого одному и тому же речевому элементу, например слову, влиянием посторонних шумов, направления и расстояния до микрофона, а также естественными вариациями характеристик голоса диктора. Ни один из известных математических методов не в состоянии компенсировать все виды изменчивости.

При решении задачи поиска ключевых слов возможны ошибки первого и второго рода, на вероятность возникновения которых влияют различные мешающие факторы, в виде акустических шумов, реверберации речевого сигнала (РС), изменения темпа произношения, состояния речевого тракта и даже время суток. Для формирования первичных признаков РС возможно использование вейвлет-пакетного разложения (ВПР), обладающего хорошей разрешающей способностью как в частотной, так и во временной областях. Кроме того, с целью придания системе устойчивости к мешающим факторам предложено использовать искусственные нейронные сети (ИНС), позволяющие обучать алгоритм в процессе эксплуатации.

Цель работы - разработка устойчивой к окружающим шумам дикторонезависимой системы автоматического поиска ключевых слов в потоке слитной речи на основе ВПР и ИНС.

Структура системы. Система поиска ключевых слов в потоке слитной речи на основе ИНС может быть разделена на четыре основных блока (рисунок 1).

Функции первичной обработки реализуются блоком ввода РС (1), состоящим из следующих узлов: микрофона, полосового фильтра, дискретизатора, схемы автоматической регулировки усиления (АРУ) и формирователя кадров.

Входной РС дискретизуется с частотой $F_д$= 8кГц и поступает на схему АРУ, которая выполнена на основе кодека с адаптивной импульсно-кодовой модуляцией (АИКМ). Из полученного после АРУ РС формируются кадры с учетом перекрытия. Затем каждый кадр подвергается ВПР и полученные отсчеты группируются в кластеры векторным квантователем. Номера

полученных кластеров и норма энергии кадра формируют вектор признаков РС. Узлы ВПР и векторный квантователь входят в блок получения первичных признаков РС (2).

Рисунок 1 – Структура системы автоматического поиска ключевых слов

Вектор признаков РС сохраняется в элементе памяти. Выделитель однородных участков устраняет инвариантность речи, связанную с изменением длительности, и разделяет входную последовательность на участки речи и паузы. Результаты работы сохраняются в элементе памяти, откуда подаются на вход ИНС обнаружителя. Элементы памяти и выделитель однородных участков входят в блок формирования слова (3). Выходной вектор ИНС поступает в решающее устройство, где окончательно производится обнаружение ключевого слова. ИНС и решающее устройство входят в блок принятия решения (4).

Основные параметры системы автоматического поиска ключевых слов на основе ИНС. Автоматическая регулировка усиления осуществляется с помощью АИКМ кодека на основе предварительного анализа параметра:

$$\sigma(n) = \alpha\sigma(n-1) + |x(n)|, \qquad (1)$$

где *x(n)* - *n* отсчет речевого сигнала, α - параметр, определяющий постоянную времени АРУ. При этом отсчет сигнала на выходе кодека АИКМ определяется как $x_0(n) = x(n)/\sigma(n)$ [3].

Использование АИКМ кодека позволит уменьшить инвариантность, связанную с изменением динамического диапазона РС.

Для получения первичных признаков реализация РС разбивалась на кадры длиной *N* с перекрытием ΔN. Исходя из максимальной надежности идентификации диктора длина кадра была выбрана равной 128 отсчетам, что соответствует 16 мс, а перекрытие кадров было выбрано 25 %. Далее, для каждого *k*-го кадра рассчитывалось ВПР.

Выбор параметров разбиения исходного РС на кадры и ВПР оказывают сильное влияние на работу системы поиска ключевых слов в целом. Также выбором этих параметров частично определяется структура ИНС. Учитывая специфику нейронных сетей, сложно предложить определенный критерий качества, который учитывал бы влияние набора параметров на надежность поиска ключевых слов. В связи с этим оптимизация параметров проводилась по критерию минимизации ошибки пропуска ключевого слова P_0 с использованием заданной структуры ИНС в виде двухслойного персептрона.

Параметры, определяющие вектор первичных признаков, могут быть разделены на две группы: параметры, характеризующие разбиение сигнала на кадры, и группа параметров, характеризующих ВПР. На длину кадра *N* накладываются ограничения, связанные с тем, что количество отсчетов сигнала в кадре должно быть равно 2^n, исходя из свойств ВПР [4]. Исходя из теории речеобразования временной интервал *m,* соответствующий кадру, не должен превышать 10...30 мс, что обусловлено квазистационарностью РС в этих пределах.

При этом размерность пространства признаков получаемого с помощью ВПР, достаточно велика, что делает задачу хранения эталонов очень ресурсоемкой. Это усугубляется еще и тем, что для обеспечения дикторонезависимости оказывается необходимым хранить по несколько эталонов на слово. Эффективным способом хранения информации, позволяющим сократить требования к памяти в десятки раз, является векторное квантование пространства признаков, когда многомерный действительный вектор признаков может быть представлен в виде номера аппроксимирующего его вектора из некоторого набора векторов, называемого кодовой книгой. Основной задачей векторного квантования является выбор кодовой книги, отражающей свойства сигнала наилучшим образом, т. е. минимизирующей ошибку аппроксимации входного сигнала векторами из кодовой книги. Для этой цели часто применяются нейронные сети, а именно их разновидность — самоорганизующиеся карты Кохонена (SOM — *self-organising map*) [2].

В векторном квантователе, построенном на нейронной сети, создается самоорганизующаяся карта Кохонена с заданной метрикой в пространстве признаков

$$d(x,y)=\sum \alpha_i \left|x_i-y_i\right|, \qquad (2)$$

где x и y — векторы свойств обучающего множества с заданным числом кластеров (нейронов второго слоя). Коэффициенты a_i задаются априорно из тех или иных соображений и учитывают вес отдельных признаков в принятии решения. Центры кластеров инициализируются случайным образом и в дальнейшем итеративно улучшаются с целью минимизации дисторсии.

Для получения номера кластера используется сеть Гроссберга [5], обучаемого из условия присвоения наибольшего номера кластеру с наименьшей вероятностью. Выход сети Гроссберга нормируется

$$NORM_n(k)=n_{кл}(k)/N_{кл}, \qquad (3)$$

где $n_{кл}(k)$ номер активного кластера *k-го* кадра, $N_{кл}$ - общее число кластеров.

Для более эффективного разделения входной последовательности на речь и паузы вычисляется норма энергии кадра

$$NORM_E(k)=E_k(k)/E_{ср}, \qquad (4)$$

где $E_к(k)$ - энергия *k-го* кадра, $E_{ср}$ - энергия РС, усредненная на интервале времени запаздывания АРУ.

Результаты работы блока формирователя признаков сохраняются в элементе памяти, представляющем собой регистр сдвига.

Для уменьшения инвариантности, связанной с изменением темпа речи, выражающейся в большей степени в изменении длительности гласных и шипящих, используется выделитель однородных участков, работа которого основана на пороговой функции:

$$W(k) \gtrless \beta W^0(k), \qquad (5)$$

где $W^0(k)=\alpha W^0(k-1)+W(k)$ - функция порога, $W(k)$ функция однородности для k кадра

$$W(k)=\left[NORM_E(k)-NORM_E(k-1)\right]^2+ \\ +\left[NORM_{кл}(k)-NORM_{кл}(k-1)\right]^2, \quad k=\overline{1,K}, \qquad (6)$$

где K – число заполненных ячеек элемента памяти, α и β коэффициенты, определяющие точность выделения однородных участков. После каждого срабатывания пороговой функции элемент памяти очищается.

Результаты работы алгоритма в виде отсчетов $NORM_n$ и $NORM_E$, соответствующих середине однородных участков, сохраняются в элементе памяти – регистре слова, на выходе которого формируется вектор признаков для работы ИНС распознавания.

При использовании ИНС одним из важных моментов является составление обучающей выборки. Подбор образцов во многом определяет характеристики работы нейронной сети [5, 6]. Используя при обучении векторы, представляющие собой сигналы, подверженные воздействию мешающих факторов, можно обеспечить робастность ИНС [7]. В рамках рассматриваемой задачи использовались образцы, на которые воздействовали узкополосные и широкополосные акустические шумы. Пары обучающих векторов состояли из входного вектора, представляющего собой набор первичных признаков РС $\vec{P}$, и выходного вектора $\vec{d}_0$, компоненты которого определяются исходя из условия $d_{0,i}=\begin{Bmatrix}1, i=N_{сл}\\ -1, i\neq N_{сл}\end{Bmatrix}$, где $N_{сл}$ - идентификатор слова, которому соответствует вектор первичных признаков РС P, подаваемый на вход ИНС. Соответствующие пары векторов были рассчитаны для исходных РС, а также для РС, подверженных воздействию мешающих факторов.

Робастные свойства, наряду с выбором обучающей выборки, во многом определяются структурой ИНС и усиливаются с увеличением числа слоев и нейронов в слоях, однако при этом значительно возрастают временные затраты на обучение, а также падает скорость работы ИНС. Большое влияние также оказывает выбор обучающей выборки и типа активационной функции [8].

Результаты экспериментальных исследований. Для экспериментального исследования была построена модель системы автоматического поиска ключевых слов. Запись эталонных реализаций, произнесенных дикторами без заметных дефектов артикуляции, осуществлялась в комнате с отсутствием акустических отражений. Эталонная совокупность признаков формировалась при нормальном темпе произношения речевых команд, обеспечивающем разборчивость речи не ниже пяти баллов в соответствии с ГОСТ Р50840-95.

Ввод и дискретизация РС осуществляются с помощью звуковой карты ЭВМ, а алгоритмы обработки выполнены в пакете MATLAB.

Для проведения исследования был собран фонетический материал в виде наборов записей 20 дикторов разного пола в возрасте от 18 до 35 лет, со средним образованием из различных рай-

онов России, записанных дважды, в разное время суток.

Каждый набор содержит 9 акустически сбалансированных фраз, регламентированных ГОСТ Р 50840-95, используемых для обучения блока формирования признаков, 20 ключевых слов, необходимых для обучения блока распознавания, и 9 тестовых фраз, содержащих ключевые слова. Записи были отсегментированы, и в результате получилось 40 наборов ключевых слов.

Одним из важных параметров ВПР является тип используемого вейвлетного фильтра[4]. В целях минимизации вычислительных затрат в данной работе рассматривались фильтры Добеши 2 порядка с глубиной разложения ВПР d=7. Таким образом, в результате экспериментального исследования были определены параметры расчета векторов первичных признаков.

Векторный квантователь представлял собой сеть встречного распространения, состоящую из двух слоев: Кохонена и Гроссберга. Слой Кохонена представлял собой однослойную нейронную сеть с линейной функцией активации. Число входов, равное 128, соответствовало числу коэффициентов ВПР. Число выходов соответствовало числу кластеров при инициализации сети и равно $N_{кл} = 500$. По завершении каждого цикла обучения, кластеры, плохо представленные в обучающем множестве и мало влияющие на дисторсию, отбрасывались. Оставшиеся кластеры использовались для получения новых значений весовых коэффициентов.

По завершении обучения, определяемого по стабилизации числа кластеров, сеть Кохонена имела 128 выходов. Слой Гроссберга имел 128 входов, соответствующих числу выходов слоя Кохонена, и 1 выход. Длина вектора первичных признаков составляла 25 кадров, что соответствовало 0,3 с. Регистр слова имел 32 ячейки, что соответствовало 16 элементам слова.

Алгоритм распознавания реализован на базе двухслойного персептрона с 32 входами, что соответствует объему элемента памяти слова. Согласно теореме Колмогорова, скрытый слой содержал 32*2+1=65 нейронов, а в выходном слое - 21 нейрон по числу идентифицируемых объектов. Использовалась биполярная сигмоидальная функция активации [8, 9].

Задача оптимизации параметров активационных функций отдельных нейронов ИНС была решена с использованием комбинированного алгоритма обучения, включающего в себя генетический алгоритм [5, 7] и модифицированный алгоритм обратного распространения ошибки [8]. Основное отличие данного метода от известного [5, 7] заключается в выборе пар векторов обучающей выборки (ОВ) не случайным образом, а по максимальной ошибке. Данный подход позволял корректировать веса синаптических связей нейронной сети не на основе случайно выбранных векторов ОВ, что требует больше итераций алгоритма, а осуществлять обучение точечно, используя образцы, на которых сеть допускает максимальную ошибку.

На рисунке 2 показана зависимость надежности распознавания от отношения сигнал-шум при действии широкополосных акустических шумов. Как следует из анализа зависимости, надежность распознавания при отношении сигнал-шум, равном 30 дБ, равна 90 %.

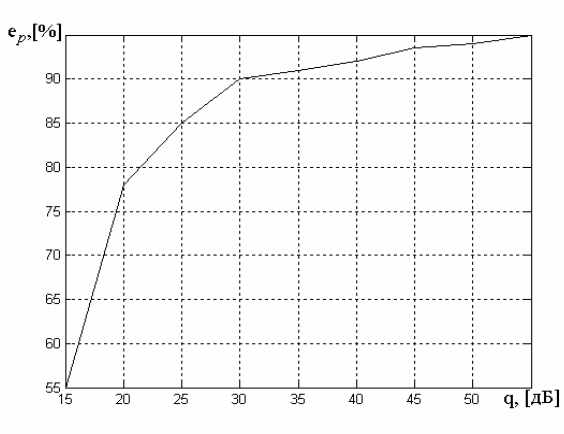

Рисунок 2 – Зависимость надежности распознавания от отношения сигнал-шум

Заключение. Уровень успешного распознавания набора из 20 слов, произнесенных 20 дикторами, при воздействии мешающих факторов, в качестве которых рассматривались акустические широкополосные шумы, в случае отношения сигнал-шум, равном 30 дБ, составлял 90 процентов.

В результате моделирования получены характеристики РС инвариантные к изменению диктора и показано уменьшение ошибки распознавания при реализации ВПР на ИНС, по сравнению с существующими системами. Получено уменьшение размерности вектора признаков, что снизило время обучения и дообучения системы. Для увеличения быстродействия системы возможна аппаратная реализация ВПР на базе ПЛИС [8, 9].

Библиографический список

1. *Галунов В.И., Соловьев Л.Н.* Темные пятна в области распознавания речи. Сборник трудов XV сессии Российского акустического общества. Т.3. М.: Геос, 2004. С. 9-19.

2. *Сорокин В.Н.* Новые концепции в автоматическом распознавании речи. Сборник трудов XIII сес-

сии Российского акустического общества. М.: Геос,1999. С. 50-57.

2. *Кириллов С.Н., Стукалов Д.Н.* Цифровые системы обработки речевых сигналов: учеб. пособие; Рязан. гос. радиотехн. акад. Рязань. 1995. 68с.

4. *Воробьев В.И., Грибунин В.Г.* Теория и практика вейвлет-преобразования. Спб.: Военный университет связи, 1999. 204 с.

5. *Дьяконов В., Круглов В.* Математические пакеты расширения MATLAB. Специальный справочник. — СПб.; Питер, 2001. — 480 с: ил.

6. *Микулич А.А.* Система распознавания набора голосовых команд на базе нейронной сети. Труды IV Международной конференции «Идентификация систем и задачи управления» SICPRO '05 Москва 2005. С. 927-933.

7. *Осовский С.* Нейронные сети для обработки информации. М.: Финансы и статистика, 2002. 344 с.

8. *Круглов А.В., Кириллов С.Н., Хахулин С.С.* Алгоритм обработки шумоподобных сигналов спутниковых систем связи на основе искусственных нейронных сетей // Электромагнитные волны и электронные системы. 2005. Т. 10. № 10. С. 27-32.

9. *Кириллов С.Н. Хахулин С.С.* Устойчивая к действию мешающих факторов система идентификации дикторов на основе искусственных нейронных сетей. Труды V Международной конференции «Идентификация систем и задачи управления» SICPRO '05 Москва 2006. С. 300-305.